Auxiliary metabolic genes in viruses infecting marine cyanobacteria

by

Luke Richard Thompson

B.S. Biological Sciences
Stanford University, 2002

Submitted to the Department of Biology
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Biology

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2010

Author .....
Department of Biology
April 14, 2010

Certified by .....
Sallie W. Chisholm
Lee and Geraldine Martin Professor of Environmental Studies
Professor of Biology
Thesis Supervisor

Certified by .....
JoAnne Stubbe
Novartis Professor of Chemistry
Professor of Biology
Thesis Supervisor

Accepted by .....
Stephen P. Bell
Professor of Biology
Investigator, Howard Hughes Medical Institute
Chairperson, Graduate Committee

Auxiliary metabolic genes in viruses infecting marine cyanobacteria

by

Luke Richard Thompson

Submitted to the Department of Biology
on April 14, 2010, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Biology

Abstract

Marine viruses shape the diversity and biogeochemical role of their microbial hosts. Cyanophages that infect the cyanobacteria *Prochlorococcus* and *Synechococcus* often carry metabolic genes not found in other bacteriophages. The proteins encoded by these 'auxiliary metabolic genes' (AMGs) are thought to increase phage fitness by altering host metabolism during infection. Dominant among the suite of AMGs carried by cyanophage are genes involved in photosynthesis and the pentose phosphate pathway (PPP). The overarching goal of this work is to understand the selective pressures driving the acquisition and maintenance of these particular AMGs by cyanophage.

Transaldolase is thought to be a key step in the host PPP. The transaldolase encoded by phage shares less than 30% amino acid identity with that of the hosts and differs in tertiary and quaternary structure despite a conserved catalytic core. The phage transaldolase was functional in vitro, and a comparison of its kinetic parameters with those of the host enzyme revealed its turnover number to be one-third that of the host. We suggest that the selection pressures underlying maintenance of the phage protein could have their origins not in kinetic properties but in genome efficiency and regulation of protein levels in the host.

Cyanophage genomes also contain genes for PPP enzymes glucose-6-phosphate dehydrogenase and 6-phosphogluconate dehydrogenase and an inhibitor of the Calvin cycle (CP12). These genes are also found in cyanophage genome fragments from the Global Ocean Sampling metagenome. Measuring their expression during infection in model phage–host pairs, we observed that phage-encoded PPP enzymes and CP12 were co-expressed with photosystem II genes while the NADPH/NADP ratio increased two-fold, consistent with increased activity of the NADPH-producing light reactions and PPP. Phage ribonucleotide reductase, which produces nucleotides using reducing equivalents from NADPH, was co-expressed with this set of genes. We propose, therefore, that phage carry these AMGs to boost the host PPP and light reactions to produce NADPH for phage genomic DNA production. No Calvin cycle AMGs have been found, supporting the hypothesis that the selection pressures molding phage genomes involve fitness advantages conferred through mobilization of host energy stores and not through carbon fixation.

Thesis Supervisor: Sallie W. Chisholm
Title: Lee and Geraldine Martin Professor of Environmental Studies
Professor of Biology

Thesis Supervisor: JoAnne Stubbe
Title: Novartis Professor of Chemistry
Professor of Biology

To my parents

Acknowledgments

The winner of the 2010 Boston Marathon finished in a record time of 2 hours, 5 minutes, and 52 seconds. Getting this PhD has been my own version of the Boston Marathon, though it took a slightly longer 59,000 hours. I have many people to thank for helping me get up Heartbreak Hill, through Kenmore Square, and past the finish line on Boylston Street.

To my thesis advisors, Penny Chisholm and JoAnne Stubbe, I thank you for your mentorship and your standard of excellence. You never let me settle for anything less than my best work, and I am a better scientist because of it. To my thesis committee, Jonathan King, Ed DeLong, and Chris Marx, thank you for your guidance, support, and thoughtful criticism throughout the thesis process. Many other teachers in my life also helped me get to this point. As an advisor, George Somero at Stanford's Hopkins Marine Station inspired in me a love of marine life, and I thank him for his continuing friendship. My first research advisor, Dennis Peters at Indiana University, welcomed me into his lab as a high school student and made time to teach me how to be a careful analytical chemist and scientific writer. Three of my teachers from North Central High School in Indianapolis also stand out. My calculus teacher Mr. Fisher and my biology teacher Mr. Russell allowed me to excel in math and science while having a lot of fun too. My literature teacher Mrs. Libby taught me to write clearly and creatively, a skill for which I will always be grateful.

To my co-workers in the Chisholm and Stubbe labs, thank you for your hard work, your ideas, and your friendship. To Debbie Lindell, Matt Sullivan, Peter Weigele, and Welkin Pope, thank you for introducing me to the wonderful world of cyanophage. To Maureen Coleman and Jake Waldbauer, thank you for your ideas and for letting the Patriots lose every now and then. To Qinglu Zeng, thank you for your hard work and your great sense of humor in the lab. To Rex Malmstrom, I will remember our good chats over dinner and great duels on the squash court. To Allison Ortigosa, Mo Seyedsayamdost, and Daniela Hristova, thank you for teaching me how to purify proteins and for welcoming me into the Stubbe lab. To Sumit Chakraborty and Yan Zhang, thanks for being there when I needed a break.

To my friends in Boston, thanks for more good times than I can remember. Maybe I could have done it without you, but it wouldn't have been as much fun. To Leonard, you've been a great friend these last seven years, and I'm really going to miss our adventures as co-rulers of the Citgoset. To Shomit, thanks for always been there, in good times and in bad times. To everyone else—Ed, Stephanie, Leslie, Dave, Israel, Anna, Annis, Pete, Tina, Jeff, Christine, Tim, Mo, Katie F., Ayman, Emilia, Michael, Hugo, Teresa, Tracy, Katie S., Elena, Jil, Craig, Maya, Adam, Kathie, Robin, Evan, Robbie, Allison, Michelle, Megan—thanks for your friendship. The adventure continues! And to Lauren, thanks for being there for me through these last few grueling months. Traveling up and down the Northeast

Corridor has never been more fun!

To my family, there is little I can say to thank you enough for your support. To Neil and Eric, thanks for being the best big brothers I could ask for. Talia and Kiva, thanks for marrying them, and giving us such an awesome niece and nephew in Nola and Ian. As if visting you all in Denver weren't fun already, now I have two more great reasons! To Lauren, thanks for supporting me. To my parents, you have given me all the tools to be a great scientist and a great person. I only hope I can live up to the example you have set.

Curriculum Vitae

Luke Richard Thompson

Born July 11, 1979, Indianapolis, Indiana

Contact Information

Permanent mailing address:
6771 E 525 S
Whitestown, IN 46075
USA

Permanent email address:
luket@alum.mit.edu

Education

Massachusetts Institute of Technology, Cambridge, Massachusetts — 2003–2010
Ph.D. Microbiology

Stanford University, Stanford, California — 1998–2002
B.S. Biological Sciences with Honors, Minor in Chemistry

North Central High School, Indianapolis, Indiana — 1994–1998
High School Diploma

Graduate Research

Biochemical and physiological role of pentose phosphate pathway genes from phages of the marine cyanobacteria *Prochlorococcus* and *Synechococcus*, Penny Chisholm and JoAnne Stubbe, advisors — 2005–2010

Genomics and metagenomics of cyanophage auxiliary metabolic genes, Penny Chisholm, advisor — 2006–2010

Prevalence of the photosystem II reaction center genes *psbA* and *psbD* in viruses infecting the marine unicellular cyanobacteria *Prochlorococcus* and *Synechococcus*, Penny Chisholm, advisor — 2004–2006

Undergraduate Research

Microsatellite analysis of intraspecific genetic variation in four species of Lake Victoria cichlids, University of Konstanz, Axel Meyer, advisor — 2000–2002

Phospholipase activity in venom of the predatory marine snail *Conus californicus*, Hopkins Marine Station of Stanford University, William Gilly, advisor — 2001

Publications

M. B. Sullivan, K. H. Huang, J. C. Ignacio-Espinoza, A. Berlin, L. Kelly, P. R. Weigele, A. S. DeFrancesco, S. E. Kern, L. R. Thompson, S. Young, W. Lee, M. Weiand, R. Fu, B. Krastins, M. Chase, D. Sarracino, M. S. Osburne, M. R. Henn, & S. W. Chisholm. Genomic analysis of oceanic cyanobacterial myoviruses compared to T4-like myoviruses from diverse hosts and environments. *Environ Microbiol*, in press, 2010.

E. R. Zinser, D. Lindell, Z. I. Johnson, M. E. Futschik, C. Steglich, M. L. Coleman, M. A. Wright, T. Rector, R. Steen, N. McNulty, L. R. Thompson, & S. W. Chisholm. Photocycle entrained patterns of global transcription, cell cycle, and photophysiology in the genetically streamlined phototroph *Prochlorococcus*. *PLoS One* 4(4):e5135, 2009.

M. Breitbart, L. R. Thompson, C. A. Suttle, & M. B. Sullivan. Exploring the vast diversity of marine viruses. *Oceanography* 20(2):135–9, 2007.

M. B. Sullivan, D. Lindell, J. A. Lee, L. R. Thompson, J. Bielawski, & S. W. Chisholm. Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biol* 4(8):e234, 2006.

Presentations

L. R. Thompson, J. Stubbe, and S. W. Chisholm. Viruses hijacking cyanobacterial carbon metabolism. Genomics: GTL Contractor-Grantee Workshop VII, Bethesda, Md. — 2009

L. R. Thompson, J. Stubbe, and S. W. Chisholm. Viruses hijacking cyanobacterial carbon metabolism. Boston Bacterial Meeting, Boston, Mass. — 2007

Awards

Poster Award, Cells, Circuits, and Computation Conference, Harvard University	2009
Genomics: GTL Conference Student Travel Grant, Department of Energy	2009
Praecis Presidential Fellowship, Massachusetts Institute of Technology	2003
Nominee, Firestone Award for Undergraduate Research, Stanford University Department of Biological Sciences	2002
Howard Hughes Summer Research Fellowship, Stanford University Department of Biological Sciences	2001

Teaching

Head Teaching Assistant, Biology 7.05. General Biochemistry, Massachusetts Institute of Technology	2007
Teaching Assistant, Biology 7.014. Introductory Biology, Massachusetts Institute of Technology	2005
Instructor, Science Explorers Program, Cambridge Community Center and Fletcher-Maynard Academy	2005–2008
Radio Talk Show Co-Host, *Biologue*, WMBR-FM Cambridge	2005–2008

Contents

List of Figures

List of Tables

Abbreviations

Pyridine nucleotides

NAD	β-nicotinamide adenine dinucleotide (oxidized form)
NADH	β-nicotinamide adenine dinucleotide (reduced form)
NADP	β-nicotinamide adenine dinucleotide phosphate (oxidized form)
NADPH	β-nicotinamide adenine dinucleotide phosphate (reduced form)

Sugar metabolites of the Calvin cycle and pentose phosphate pathway

BPG	2,3-bisphosphoglycerate
DHAP	dihydroxyacetone phosphate
E4P	erythrose 4-phosphate
FBP	fructose 1,6-bisphosphate
F6P	fructose 6-phosphate
GAP	glyceraldehyde 3-phosphate
G6P	glucose 6-phosphate
PGA	3-phosphoglyceric acid
R5P	ribose 5-phosphate
RuBP	ribulose 1,5-bisphosphate
Ru5P	ribulose 5-phosphate
SBP	sedoheptulose 1,7-bisphosphate
6PG	6-phosphogluconate
6PGL	6-phosphogluconolactone
S7P	sedoheptulose 7-phosphate
X5P	xylulose 5-phosphate

Genes and proteins of the Calvin cycle and pentose phosphate pathway

cbbA[1]	aldolase	fructose-1,6-bisphosphate aldolase
cbbA[1]	aldolase	sedoheptulose-1,7-bisphosphate aldolase
cp12	CP12	Calvin cycle inhibitor CP12
gap2	GAPDH	glyceraldehyde-3-phosphate dehydrogenase
glpX[2]	FBPase	fructose-1,6-bisphosphatase
glpX[2]	SBPase	sedoheptulose-1,7-bisphosphatase
gnd	6PGDH	6-phosphogluconate dehydrogenase
opcA	OpcA	glucose-6-phosphate dehydrogenase effector OpcA
pgi	PGI	phosphoglucose isomerase
pgk	PGK	phosphoglycerate kinase
pgl	6PGLase	6-phosphogluconolactonase
prkB	PRK	phosphoribulokinase
rbcLS	rubisco	ribulose-1,5-bisphosphate carboxylase/oxygenase
rpe	RPE	ribulose-5-phosphate epimerase
rpiA	RPI	ribulose-5-phosphate isomerase
tal	TA	transaldolase
tktA	TK	transketolase
tpi	TPI	triosephosphate isomerase
zwf	G6PDH	glucose-6-phosphate dehydrogenase

[1]In all organisms, *cbbA* encodes a dual-functional FBP aldolase/SBP aldolase.

[2]In most cyanobacteria and proteobacteria, *glpX* encodes a dual-functional FBPase/SBPase.

Nomenclature

Prochlorococcus, *Synechococcus*, and their phages are named according to a defined nomenclature, with some exceptions. We do not define species in either the cyanobacteria or their phages. Rather, we refer to them at the level of genera, defining genetically unique strains within those genera. *Prochlorococcus* and marine *Synechococcus* are closely related sister taxa; freshwater *Synechococcus* is more distantly related to the marine variety and *Prochlorococcus*, yet regrettably it shares the same name. Unless otherwise noted, all references to *Synechococcus* in this thesis are to marine *Synechococcus*. Cyanophages are named according to the strain on which they were isolated—*Prochlorococcus* or *Synechococcus*—although some phages are able to infect both genera.

Prochlorococcus and *Synechococcus*

Strains are named with initials of the institution at which they were isolated and originally maintained, followed by the last two digits of the year of isolation, and finally the isolate number. The strain name is preceded by the genus name "*Prochlorococcus*" or "*Synechococcus*" in the first mention, but the genus is usually dropped in further references. In general, "MIT" strains are *Prochlorococcus*, and "WH" strains are *Synechococcus*. Some examples are shown below.

MIT9312	*Prochlorococcus* isolated by MIT in 1993, isolate no. 12
WH8102	*Synechococcus* isolated by Woods Hole in 1981, isolate no. 2
MED4	*Prochlorococcus* isolated from the Mediterranean Sea, isolate no. 4 (does not follow convention)
SS120	*Prochlorococcus* isolated from the Sargasso Sea, isolated from 120 meters (does not follow convention)

Cyanophages

Phage strains are named with the initial of the genus on which they were isolated (P, *Prochlorococcus*; S, *Synechococcus*), followed by a dash, followed by the location from which they were isolated, then morphology type (P, podovirus; M, myovirus; S, siphovirus), then isolate number. Because some locations have two initials and others just one, it is easiest to read the last set of letters from the right. Some examples are shown below.

P-HM1	*Prochlorococcus* phage from Hawai'i, myovirus, isolate no. 1
P-SSM4	*Prochlorococcus* phage from the Sargasso Sea, myovirus, isolate no. 4
P-SSP7	*Prochlorococcus* phage from the Sargasso Sea, podovirus, isolate no. 7
P-SS2	*Prochlorococcus* phage from slope waters, siphovirus, isolate no. 2
S-RSM2	*Synechococcus* phage from the Red Sea, myovirus, isolate no. 2
Syn5	*Synechococcus* phage, isolate no. 5 (does not follow convention)
Syn9	*Synechococcus* phage, isolate no. 9 (does not follow convention)

CHAPTER 1

Introduction to auxiliary metabolic genes in cyanophage

Introduction

The co-evolution of viruses and their hosts has shaped nearly all life on Earth, from bacteria to humans. The exchange of genetic material between viruses and their hosts leaves in its wake significantly altered viral and host genomes. The human genome, for example, is approximately 8% viral DNA, in the form of retroviral proviruses (Horie et al. 2010), and human retroviruses carry host-derived oncogenes that lead to cancer (Maeda et al. 2008). In the prokaryotic realm, some 13% of genes found only in cyanobacteria bear signs of horizontal gene transfer (HGT) within cyanobacteria (Yerrapragada et al. 2009), and phage-mediated HGT is thought to be a major mechanism of HGT in cyanobacteria (Zeidner et al. 2005). Bacteriophage encode a tremendous diversity of host-derived genes, from nucleotide biosynthesis genes (Chen et al. 1995) to ribosomal RNA at the heart of bacterial protein synthesis (Beumer and Robinson 2005).

The use of host-like metabolic genes by viruses is a central theme in the co-evolution of viruses and their hosts. The advent of DNA sequencing and its application to the compact genomes of bacteriophages (Sanger et al. 1977) and bacteria (Fleischmann et al. 1995) has led to the discovery of remarkable gene acquisitions by phage. The sequencing of cyanophage genomes (Chen and Lu 2002, Sullivan et al. 2005, Pope et al. 2007, Weigele et al. 2007, Millard et al. 2009, Sullivan et al. 2009, Henn et al. 2010) in particular has identified genes from core metabolic pathways in their host cyanobacteria. These metabolic genes provide clues to the mechanisms of cyanophage replication, betraying a tight association between the cyanophage life cycle and cyanobacterial physiology. Metabolic genes in viruses and their roles during infection are the focus of this thesis. In this introductory chapter, I consider first the players: the marine cyanobacteria *Prochlorococcus* and *Synechococcus*

and the viruses (cyanophage) that infect them. I then consider the genes these viruses are known to carry: what metabolic pathways the products of these genes appear to be involved in and how they might function in these pathways.

Marine cyanobacteria and cyanophage

The marine cyanobacteria *Prochlorococcus* and *Synechococcus*, which are the numerically most abundant photosynthetic organisms on the planet (Partensky et al. 1999, Scanlan and West 2002), are a rich source of energy and biomass for viruses. In oligotrophic waters between 40°N and 40°S, *Prochlorococcus* can regularly achieve densities of 10^4–10^5 mL^{-1} (DuRand et al. 2001, Zinser et al. 2006, Johnson et al. 2006). In coastal and more mesotrophic and eutrophic waters, *Synechococcus* often dominates (Zwirglmaier et al. 2008), but its densities are more variable. Global abundance of *Prochlorococcus* has been estimated at 10^{27} cells (Scanlan et al. 2009). *Prochlorococcus* cells fix an estimated 1.5×10^{10} kg of carbon per day and constitute a global biomass of 1.2×10^{11} kg (Garcia-Pichel et al. 2003). Similar estimates of *Synechococcus* abundance, carbon fixation, and biomass are about one-half to one-third of these estimates, with the *Synechococcus* global biomass estimated at 4.3×10^{10} kg (Garcia-Pichel et al. 2003).

Collectively, marine viruses often outnumber cyanobacteria and other bacteria by an order of magnitude or more, often found at densities of 10^7 mL^{-1} (Bergh et al. 1989). There are an estimated 10^{31} viruses in the global ocean, which constitute a mass of 4×10^{11} kg (Suttle 2005). This is more than the mass of all living humans (3×10^{11} kg). It is unknown how many of these viruses are specific for *Prochlorococcus* or *Synechococcus*, but viruses co-occuring with these cyanobacteria are regularly isolated on cultured strains (Waterbury and Valois 1993). Cyanophages titered on specific *Prochlorococcus* and *Synechococcus* strains in culture have been found at lower densities (10^3 mL^{-1}) than their hosts (Sullivan et al. 2003). This titer is four orders of magnitude lower than the concentration of total viruses, and it likely reflects the limited spectrum of available host strains as well as an incomplete understanding of phage–host dynamics. The cyanophages that have been isolated on *Prochlorococcus* and *Synechococcus* all fall into three types based on morphology, life cycle, and genome structure: T4-like myoviruses, T7-like podoviruses, and siphoviruses (Sullivan et al. 2003) (Figure 1-1). Among this set, T4-like myoviruses are the most commonly

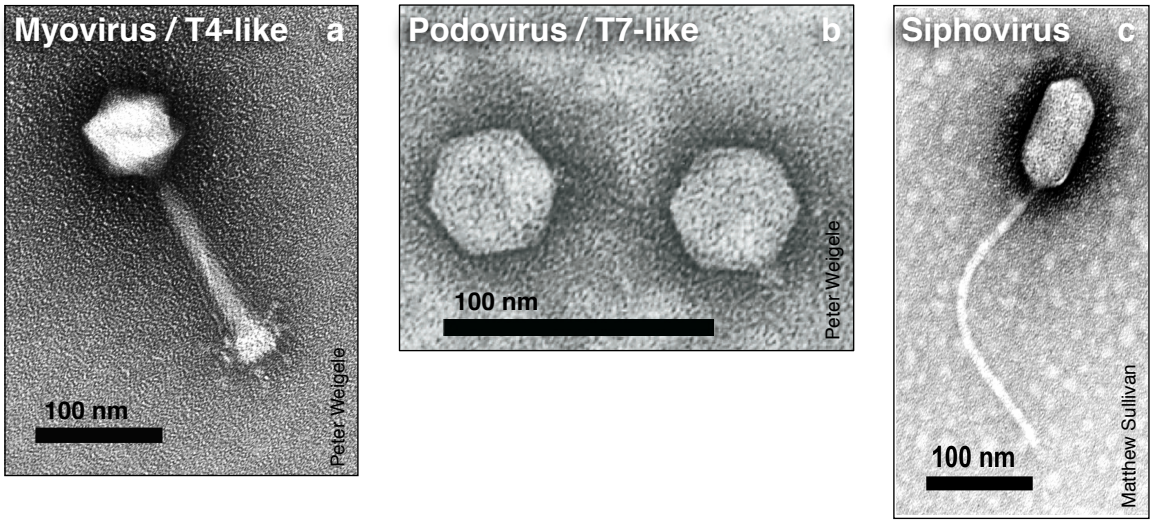


Figure 1-1: Electron micrographs of cyanophages. Shown are (a) T4-like myovirus P-SSM4 (Sullivan et al. 2005), (b) T7-like podovirus P-SSP7 (Sullivan et al. 2005), and (c) siphovirus P-SS2 (Sullivan et al. 2009).

isolated from the environment, followed by T7-like podoviruses and then siphoviruses (Sullivan et al. 2003). This pattern is presumed to reflect the relative abundances of these phage types in the ocean, although it is admittedly restricted by the host strains used to isolate the viruses.

Cyanophage may be important agents of host mortality (Suttle and Chan 1994), creating selective pressures on their hosts (Sandaa et al. 2009) and feeding the microbial food web (Brussaard et al. 2008). The exact extent of their role in host mortality has been called into question, however, because of evidence suggesting that most cyanobacteria co-occurring with cyanophage are resistant to infection (Waterbury and Valois 1993). Other studies have shown that as many as 12% (DeLong et al. 2006) to 14% (Suttle and Chan 1994) of marine cyanobacteria are infected at any given time.

Cyanophage are agents of horizontal gene transfer (HGT), moving genes among their hosts. Evidence for cyanophage-mediated HGT is mostly indirect: host genomic islands with phage genes (Coleman et al. 2006) and cyanophages carrying genes with high sequence identity to host genes (Mann et al. 2003, Sullivan et al. 2006). The acquisition of genes from the host is of particular interest because it can confer new metabolic capabilities to cyanophages. By carrying host genes, cyanophages can also act as reservoirs of host genetic

diversity (Sullivan et al. 2006).

A useful model for cyanophage infection is the *E. coli* bacteriophage T4, as it is a well-characterized model phage and, as mentioned above, T4-like phage are the most commonly isolated type of cyanophage (Sullivan et al. 2003), with their genes frequently observed in marine metagenomic databases (Williamson et al. 2008). Common features of the T4-like phage life cycle are shown in Figure 1-2. In addition to phage-encoded enzymes that take over the host replication and protein-synthesis machinery, T4 makes fundamental changes to metabolism of the host cell. Phage-encoded enzymes dephosphorylate (Depew and Cozzarelli 1974), methylate (Mathews and Kessin 1967), phosphorylate (Sakiyama and Buchanan 1971), and glucosylate (Gold and Schweiger 1969) nucleotides, cleave DNA (Yasuda and Sekiguchi 1970), polymerize DNA (Waard et al. 1965), and ligate DNA and RNA (Fareed and Richardson 1967). As these enzymatic activities suggest, however, most of the manipulations of host metabolism by T4 involve DNA metabolism. Beyond this rich toolkit of genes for degrading and synthesizing individual nucleotides and larger pieces of DNA, there are no 'host genes' for altering host metabolism—for example, central carbon metabolism—in the T4 genome (Miller et al. 2003b). Perhaps because there are ample energy, energy storage, and building blocks available in *E. coli*, T4 does not carry genes for enzymes for many of the core metabolic pathways of its host. As we will see next, the situation is strikingly different in T4-like cyanophage and other types of cyanophage.

Auxiliary metabolic genes in cyanophage

Phage genes with putative functions in host metabolism have been termed 'auxiliary metabolic genes' (AMGs) (Breitbart et al. 2007, Appendix E). Known cyanophage AMGs encode proteins involved in the light reactions of photosynthesis (Mann et al. 2003), the pentose phosphate pathway (PPP) (Millard et al. 2004), phosphate acquisition (Sullivan et al. 2005), and DNA biosynthesis (Sullivan et al. 2005). These phage-encoded proteins are thought to play a role in host metabolism during infection, leading to a more productive infection (i.e., more phage progeny). The fact that many of the AMGs found in cyanophage, such as those for photosynthesis, lie at the heart of the energy-generation process of cyanobacteria hints that, for cyanophage, and perhaps for phage in general, the maintenance of core elements of host metabolism during infection is important.

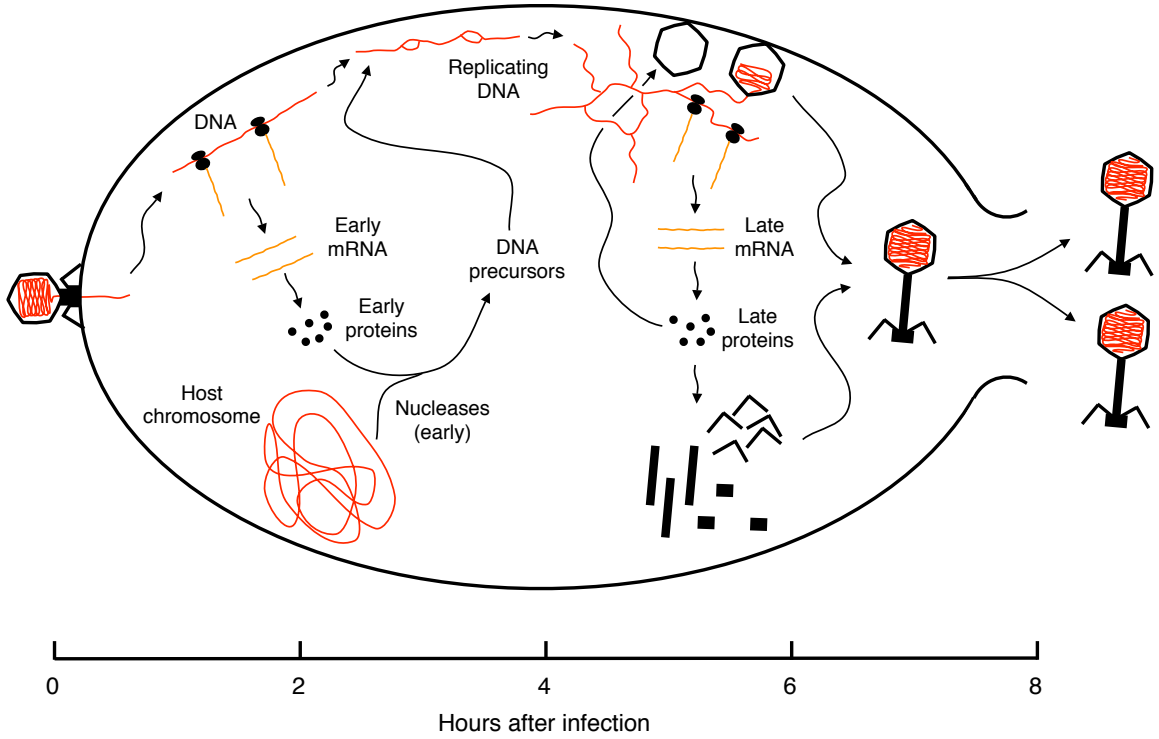


Figure 1-2: Life cycle of a T4-like phage. Following translocation of phage DNA into the cell, phage replication proceeds along a defined course. In T4-like cyanophage, the lytic cycle takes ∼8 h (see Chapters 3 and 4), whereas in T4, infecting *E. coli*, it is much more rapid, lasting only 25 min (Mathews 1977, Hadas et al. 1997); this likely reflects the short generation time of *E. coli* (20 min) relative to that of cyanobacteria (1 day). DNA is transcribed by the host RNA polymerase (early genes) and then by a phage-modified host RNA polymerase (late genes). Early genes are involved in establishing infection and encode enzymes or regulatory proteins; these include nucleases to degrade the host chromosome, various enzymes to synthesize DNA building blocks, and polymerases to replicate the phage genome (Mathews 1977, Miller et al. 2003b, Roucourt and Lavigne 2009). Late genes are involved in assembling progeny virions and encode structural proteins; these include capsid and tail-fiber proteins (Mathews 1977, Miller et al. 2003b, Roucourt and Lavigne 2009).

Two criteria are used to distinguish AMGs from other genes in phage genomes. A phage gene is an AMG if it is

1. putatively involved in host metabolism rather than in a phage-specific process and
2. not generally found in phages from other types of hosts and environments.

By the first criterion, a structural gene or a gene found only in phage would not qualify. By the second criterion, a metabolic gene that has a host homolog but is found in phages of multiple hosts, from multiple environments—e.g., in both cyanophage and coliphage—would not qualify. Ribonucleotide reductase (RNR) genes are difficult to classify, for example, because they appear to be enriched in cyanophage, being found in all sequenced *Prochlorococcus* and *Synechococcus* cyanophages (Chen and Lu 2002, Sullivan et al. 2005, Pope et al. 2007, Weigele et al. 2007, Millard et al. 2009, Sullivan et al. 2009), yet are also found in some non-cyanophages, e.g., non-cyanophage T4-like phages (Petrov et al. 2006, Comeau et al. 2007). Photosynthesis genes in cyanophage genomes, however, are clear instances of AMGs because they are clear homologs of host-specific metabolic genes and are not found in non-cyanophage (Sullivan et al. 2006). In spite of occasional difficulties in classifying certain phage genes as AMGs, the term nevertheless serves a purpose because it provides language that helps us distinguish host- and environment-specific phage genes from common, 'housekeeping' phage genes.

Because phage have no metabolism of their own and must tap into host metabolic pathways, the AMGs carried by cyanophage are tied to the metabolism and lifestyle of marine cyanobacteria. Cyanobacteria are unique among bacteria in that they obtain energy from oxygenic photosynthesis. As a result, they use genes and pathways not found in non-cyanobacteria, and their physiology is fundamentally different from non-cyanobacteria. Phages infecting cyanobacteria therefore have access to a unique gene set, and they are subject to the unique selective pressures of infecting a photosynthetic host, so it is perhaps not surprising that cyanophage carry a unique complement of genes. The marine realm has unique properties relative to the habitat of various non-marine bacteria, and this places additional pressures on host physiology. For example, many parts of the ocean are phosphate-limited (Wu et al. 2000, Thingstad et al. 2005). Therefore it is not surprising to find genes for coping with phosphate stress in cyanophage genomes. This is clearly a function of the marine environment rather than just cyanobacterial physiology, as phosphate-stress genes

are also found in phages of marine *Vibrio* (Miller et al. 2003a).

Because AMGs are host- and environment-specific, we postulate that they have been acquired and maintained to fill key host-specific metabolic bottlenecks during infection. If the reaction for converting substrate (S) to product (P), catalyzed by the host enzyme (E_{host}), becomes limiting during infection, and if cyanophage replication depends on having enough P, the phage may encode its own enzyme (E_{phage}) to prevent this reaction from becoming limiting. This reaction could be limiting because there is insufficient E_{host}, and therefore E_{phage} increases the amount of total enzyme, or because the activity of E_{host} is suppressed under certain conditions, and E_{phage} is active under those same conditions. Importantly, this reaction may not necessarily be a bottleneck during normal host growth. More likely, P becomes limiting during and because of the infection process. For example, infection places a high demand on DNA biosynthesis (Paul et al. 2002), which the phage 'directs' to produce nucleotides for phage genome replication, likely decoupled from DNA biosynthesis and replication processes in the host (Clokie and Mann 2006). This biosynthesis requires both reduced carbon and energy, which can be produced only by cyanobacterial metabolism. As described below, genes involved in several pathways that make precursors for nucleotide biosynthesis have been found in cyanophage genomes (Table 1.1). These appear to be cyanophage/cyanobacteria-specific pathways for synthesizing DNA building blocks, and as we will argue in the following chapters, the particular genes encoded by cyanophage for these pathways may represent the key bottlenecks in these pathways during infection.

Metabolic pathways represented by AMGs

Photosynthesis

No host metabolic pathway is more represented in cyanophage genomes than photosynthesis (Table 1.1 and Figure 1-3). Several categories of photosynthesis genes are represented. Cyanophage genomes carry genes (*psbA*, *psbD*) for photosystem II (Mann et al. 2003, Sullivan et al. 2006), and D1 protein synthesized from cyanophage *psbA* has been detected during infection (Lindell et al. 2005). Additionally, photosystem I gene cassettes have been found in environmental DNA of likely cyanophage origin (Sharon et al. 2009). Other cyanophage genes in photosynthetic electron transport include plastoquinol terminal oxidase (PTOX),

Table 1.1: Partial list of auxiliary metabolic genes in cyanophages. A gene is marked with '×' if at least one phage genome in that group has this gene; '(×)' signifies that this gene has been found only in putative phage metagenomic sequence; '[×]' signifies that the presence of this gene is reported here for the first time. Abbreviations: T4-likes, T4-like myoviruses; T7-likes, T7-like podoviruses; siphos, siphoviruses; AICARFT/IMPCHase, phosphoribosylaminoimidazolecarboxamide formyltransferase/IMP cyclohydrolase; AIR synthetase, phosphoribosylaminoimidazole synthetase; GAR transformylase, phosphoribosylglycinamide formyltransferase; FGAM synthetase, phosphoribosylformylglycinamidine synthetase; OPRTase, orotate phosphoribosyltransferase. References: Chen and Lu (2002), Sullivan et al. (2005), Pope et al. (2007), Weigele et al. (2007), Millard et al. (2009), Sullivan et al. (2009), Sharon et al. (2009).

Pathway	Gene	Function	T4-likes	T7-likes	Siphos
Photosynthesis					
	psbA	photosystem II core protein D1	×	×	
	psbD	photosystem II core protein D2	×		
	PTOX	plastoquinol terminal oxidase	×		
	petE	plastocyanin	×		
	psaA	photosystem I P700 chlorophyll a apoprotein A1	(×)		
	psaB	photosystem I P700 chlorophyll a apoprotein A2	(×)		
	psaC	photosystem I iron–sulfur center	(×)		
	psaD	photosystem I reaction center subunit II	(×)		
	psaE	photosystem I reaction center subunit IV	(×)		
	psaK	photosystem I reaction center subunit	(×)		
	psaJF	photosystem I reaction center subunit III/IX	(×)		
	petF	ferredoxin [2Fe–2S]	×		
	hli	high-light-inducible protein	×	×	
	ho1	heme oxygenase	×		
	pebS	phycoerythrobilin synthase	×		
	pcyA	phycocyanobilin:ferredoxin oxidoreductase	×		
	cpeT	phycoerythrin biosynthesis	×		
Carbon metabolism					
	talC	transaldolase	×	×	
	cp12	Calvin cycle inhibitor CP12	[×]	[×]	[×]
	zwf	glucose-6-phosphate dehydrogenase	×		
	gnd	6-phosphogluconate dehydrogenase	×		
Phosphate acquisition					
	phoH	unknown phosphate-stress-induced protein	×		
	pstS	ABC-type phosphate transport system	×		
	phoA	alkaline phosphatase	×		
Nucleotide biosynthesis					
	nrdA	ribonucleotide reductase, class I, alpha subunit	×		
	nrdB	ribonucleotide reductase, class I, beta subunit	×		
	nrdJ	ribonucleotide reductase, class II		×	×
	cobS	cobalt chelatase (cobalamin biosynthesis)	×		
	purH	bifunctional AICARFT/IMPCHase (purine biosynthesis)	×		
	purM	AIR synthetase (purine biosynthesis)	×		
	purN	GAR transformylase (purine biosynthesis)	×		
	purS	FGAM synthetase (purine biosynthesis)	×		
	pyrE	OPRTase (pyrimidine biosynthesis)	×		
	thyX	thymidylate synthase (pyrimidine biosynthesis)	×	×	×

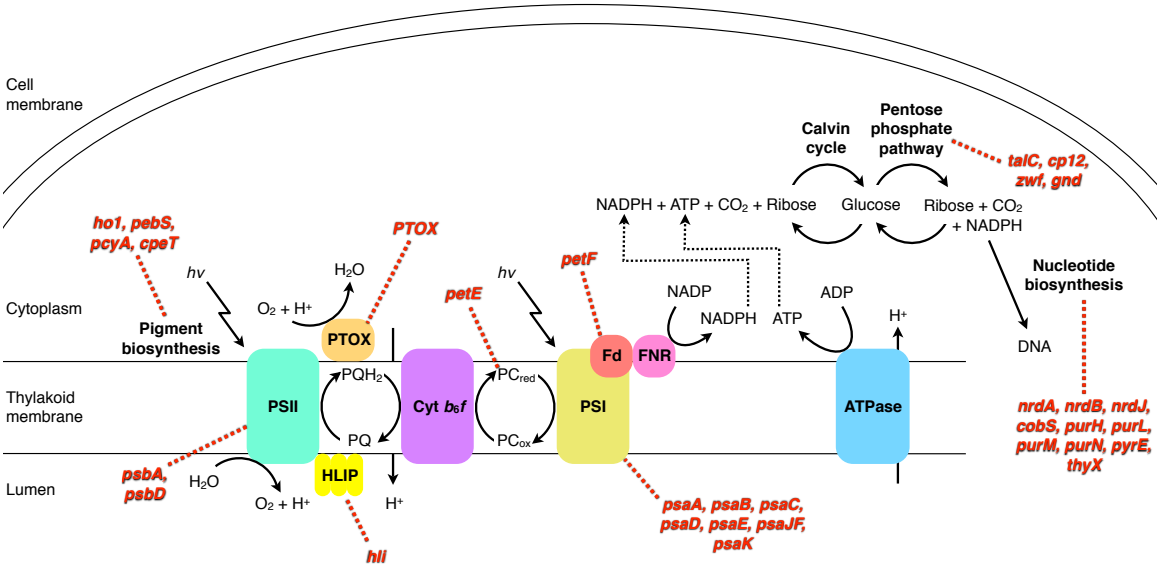


Figure 1-3: Locations of cyanophage AMGs in cyanobacteria photosynthesis and carbon metabolism. Cyanophage genes are marked in red. Abbreviations: PSII, photosystem II; PQ, plastoquinone (oxidized); PQH_2, plastoquinone (reduced); PTOX, plastoquinol terminal oxidase; cyt b_6f, cytochrome b_6f complex; PC_{ox}, plastocyanin (oxidized); PC_{red}, plastocyanin (reduced); PSI, photosystem I; Fd, ferredoxin; FNR, ferredoxin–NADP reductase; PPP, pentose phosphate pathway.

plastocyanin (*petE*), and ferredoxin (*petF*) (Sullivan et al. 2005, Weigele et al. 2007, Millard et al. 2009). Cyanophage also carry genes for the biosynthesis of several photosynthetic pigments, including phycobilin (*ho1*, *pebS*, *pcyA*) and phycoerythrin (*cpeT*) (Sullivan et al. 2005, Weigele et al. 2007, Sullivan et al. ress). The phycobilin biosynthesis genes have been shown to be functional when heterologously expressed in *E. coli* (Dammeyer et al. 2008), and the phage ferredoxin is a reductant in vitro for the PebS-mediated transformation (Dammeyer et al. 2008), a concerted reaction that requires two separate enzymes in the host. Cyanophage genomes are also rich in genes for high-light-inducible proteins (*hli*) (Lindell et al. 2004), which have been proposed to function to protect the photosynthetic machinery from excess radiation or in a more general stress response (He et al. 2001). Notably absent from the set of photosynthesis genes in cyanophage are genes for the Calvin cycle. No Calvin cycle genes have been identified in cyanophage genomes or in metagenomics samples from the natural environment (see Chapter 3).

The proposed role of photosynthesis genes during infection is to maintain photosynthesis

to produce energy and to avoid significant photodamage and reactive oxygen species (ROS) formation (Mann et al. 2003, Lindell et al. 2005). Phage D1 in particular is implicated because synthesis of host D1 in photosystem repair is inhibited by ROS (singlet oxygen and hydrogen peroxide) formed as a result of excess photosynthetic electron flow (Latifi et al. 2009). Nearly all stages (i.e., photosystem II, plastocyanin pool, photosystem I, etc.) of the photosynthetic electron transport chain are represented in at least one cyanophage genome (Sullivan et al. 2005, Weigele et al. 2007, Millard et al. 2009) or metagenomic fragment (Sharon et al. 2009) (Table 1.1 and Figure 1-3). The presence of PTOX in some cyanophages, which dissipates photosynthetic electron flow without producing energy (Bailey et al. 2008), suggests that at least part of the purpose of these genes is to maintain electron flow in a photoprotective role. However, the presence of PPP genes for producing energy as reducing equivalents—as previously reported (Millard et al. 2004, Sullivan et al. 2005, Weigele et al. 2007) and described below—indicates that energy generation is important for phage replication, and photosynthesis likely plays a role here too. Because they encode photosynthesis genes, cyanophage have been proposed to contribute to global primary production. Sharon et al. (2007), citing the high proportion (60%) of total *psbA* genes in the open ocean that are of cyanophage origin, have argued, "Phage-encoded proteins may play a direct role in determining the level of photosynthetic productivity in oceans (oxygen evolution and carbon fixation)." We find this argument lacking, however, as phage-infected cells are dying, and even a temporary increase in photosynthetic activity would be more than offset by the loss in production from the cells (and their future offspring) upon their death.

Carbon metabolism

Cyanophage carry several genes for the PPP (Table 1.1 and Figure 1-3) (Millard et al. 2004, Sullivan et al. 2005, Weigele et al. 2007). The PPP oxidizes glucose to make NADPH and ribose, with carbon dioxide as a by-product; notably, this is the opposite of the Calvin cycle, which uses NADPH (along with ATP) to fix carbon dioxide into glucose (Stanier and Cohen-Bazire 1977) (Figure 1-3). Specifically, there are three genes involved in the PPP carried by cyanophage. These include the NADPH-producing enzymes glucose-6-phosphate dehydrogenase (*zwf*) and 6-phosphogluconate dehydrogenase (*gnd*) (Weigele et al. 2007) and the sugar transferase transaldolase (*talC*) (Millard et al. 2004, Sullivan et al. 2005).

Although some PPP enzymes are shared with the Calvin cycle, all three of these (*zwf*, *gnd*, *talC*) are exclusive to the PPP (see Chapter 3).

As will be described in Chapter 3, many cyanophages also carry a gene for the Calvin cycle inhibitor CP12 (*cp12*), which in other cyanobacteria has been shown to inhibit the Calvin cycle enzymes phosphoribulokinase and glyceraldehyde-3-phosphate dehydrogenase (Tamoi et al. 2005). This serendipitous discovery arose from my study of the regulation of carbon metabolism in *Prochlorococcus* MED4 over the light–dark cycle (Zinser et al. 2009, Appendix F). There we reported the presence of *cp12* in *Prochlorococcus* genomes, which in turn led me to look for and find *cp12* in cyanophage genomes (Sullivan et al. in press, Appendix G). This finding has provided key insights into the regulation of host carbon metabolism by cyanophage.

The proposed role of PPP genes during infection is to produce NADPH and ribose from stored carbon, either in the dark (Millard et al. 2004) or possibly irrespective of the light–dark cycle (Sullivan et al. 2005). Importantly, both NADPH and ribose are precursors of DNA biosynthesis. Significant discussion in the following chapters will be devoted to the potential role of cyanophage AMGs in the host PPP, with an emphasis on their possible relationship to nucleotide biosynthesis.

Phosphate acquisition

Several genes for phosphate acquisition are found in cyanophage genomes (Table 1.1 and Figure 1-3) (Sullivan et al. 2005, Weigele et al. 2007, Millard et al. 2009). The most prevalent of these, *phoH*, has an ATPase domain (Kim et al. 1993) but no known function, but it is part of the phosphate regulon in *E. coli* and is up-regulated in response to phosphate stress in that organism (Kim et al. 1993). Curiously, *phoH* is not induced under phosphate starvation in *Prochlorococcus* (Martiny et al. 2006). *phoH* is found in all T4-like cyanophages, a marine T4-like vibriophage, and a marine T7-like roseophage, but not in any non-marine phages (Sullivan et al. in press, Appendix G). Thus, although *phoH* is not unique to cyanophages, it is unique to marine phages, suggesting its function is tied to selective pressures of the marine environment (i.e., low phosphate). A high-affinity phosphate transporter (*pstS*) has also been found in several cyanophage genomes (Sullivan et al. 2005). Notably, of the cyanophages in which *pstS* is found, all come from oligotrophic waters, where phosphate concentrations are likely low (Sullivan et al. in press, Appendix G). Some cyanophages also

encode alkaline phosphatase (*phoA*), which has been proposed to provide access to organic phosphorus, either from the environment or from within the host (Sullivan et al. in press, Appendix G).

The proposed role of phosphate acquisition genes during infection is to scavenge phosphate, which may be limiting in low-phosphate environments (Sullivan et al. 2005). One likely use of phosphate during infection is nucleic acid biosynthesis. Specifically, as discussed, phage replication places a high demand on DNA biosynthesis for phage genome replication, and we would expect phosphate to be an important and possibly limiting substrate for nucleotide biosynthesis. Particularly given the oligotrophic and phosphate-deplete conditions of many *Prochlorococcus* and *Synechococcus* populations, it is likely that phosphate limitation places significant strain on this pathway. Phage-encoded phosphate transport and acquisition systems could help alleviate this potential bottleneck.

Nucleotide biosynthesis

Finally, nucleotide biosynthesis genes themselves are commonly found in cyanophage genomes (Table 1.1 and Figure 1-3). The most common of these is ribonucleotide reductase (*nrdAB/J*) (Sullivan et al. 2005, Weigele et al. 2007, Millard et al. 2009). Based on their prevalence in cyanophage, RNR enzymes are likely critical for cyanophage to convert ribonucleotides to deoxyribonucleotides. Related to RNR, many T4-like cyanophages encode a cobalt chelatase subunit (*cobS*) (Sullivan et al. 2005), which combined with two other subunits (*cobN*, *cobT*) inserts cobalt into cobalamin (B_{12}) (Rodionov et al. 2003), a critical cofactor for the host class II RNR (Stubbe et al. 2001). These T4-like cyanophages encode their own class I RNR (Sullivan et al. 2005), which has no requirement for cobalamin (Stubbe et al. 2001), and thus it appears they exploit both their own RNR and the host RNR for DNA biosynthesis. Cyanophage also encode several genes for purine biosynthesis (*purH*, *purL*, *purM*, *purN*) and pyrimidine biosynthesis (*pyrE*, *thyX*). Most of these are found sporadically in T4-like cyanophages, but thymidylate synthase (*thyX*) is notable for its presence in all three types of cyanophage.

The proposed role of DNA biosynthesis genes during infection is, clearly, to synthesize nucleotides for phage genome replication (Klumpp et al. 2008, Alemayehu et al. 2009). With help from various purine and pyrimidine biosynthetic enzymes, and most importantly RNR to convert RNA nucleotides to DNA nucleotides, cyanophage catalyze many of the steps

to make their own DNA. This is important, as degradation and reuse of DNA from the host chromosome is potentially limited by the genome size of the host (Paul et al. 2002), and *Prochlorococcus* has the smallest genome of any cyanobacterium (Rocap et al. 2003, Kettler et al. 2007). De novo synthesis of DNA building blocks is likely critical for infecting cyanophage to achieve optimal burst sizes, and there is thought to be strong selection for high burst size in lytic phages living in dilute environments (Clokie and Mann 2006).

Hypothesized role of AMGs during infection

Putting all these pieces together, we can imagine a scenario in which phage-encoded metabolic enzymes both 'take over' and 'tune' the existing host metabolism to provide energy and biomass (carbon skeletons) for phage replication, with a particular emphasis on replication of the phage genome. Figure 1-4 shows how we might conceptualize the arrangement of host metabolism to produce DNA nucleotides for its own genome replication (pane a) and how we might expect this metabolism to be altered upon phage infection (pane b). In uninfected *Prochlorococcus* (Figure 1-4a), daytime photosynthetic electron transport produces ATP and NAPDH, which power carbon fixation in the Calvin cycle, also during the day. At night, the PPP uses this glucose to generate NADPH and ribose, which are used along with phosphate transported from the environment to synthesize DNA nucleotides and *Prochlorococcus* genomic DNA. In infected *Prochlorococcus* (Figure 1-4b), the situation is similar but with one significant difference: energy from photosynthetic electron transport is not used by the Calvin cycle because it is inhibited by phage-encoded CP12. Rather, ATP and NADPH from the light reactions combine with NADPH from the PPP to power nucleotide biosynthesis. The light reactions, the PPP, phosphate acquisition, and nucleotide biosynthesis are all boosted by phage-encoded enzymes, whereas the Calvin cycle is shut off by phage-encoded CP12.

This is a model that is consistent with known information, but it raises many questions that require experimental validation. Most of the discussion thus far has focused on gene and genome sequences, which give no guarantee that genes are functional. *Are the proteins encoded by AMGs functional in vitro, and are there differences between phage and host versions of orthologous proteins?* This is particularly relevant to the phage transaldolase (TalC), which has less than 30% amino acid identity with the host transaldolase (TalB), in-

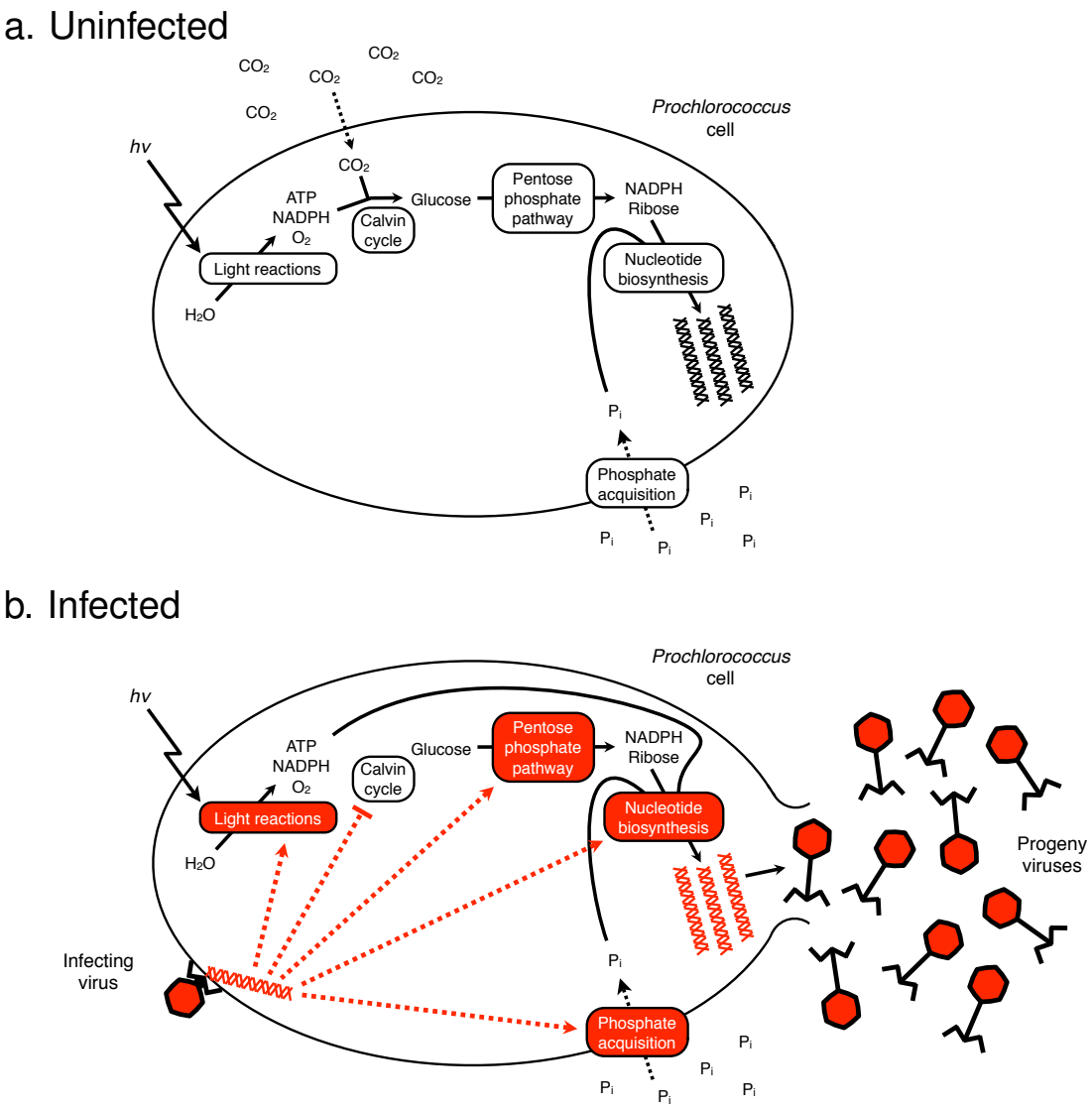


Figure 1-4: Schematic of metabolism in uninfected and infected *Prochlorococcus*. (a) Uninfected *Prochlorococcus* metabolism converts inorganic starting materials (carbon dioxide, water, light, phosphate, and nitrogen) to biomass. Focusing on the biosynthesis of DNA, this transformation encompasses five fundamental metabolic pathways: the light reactions of photosynthesis, the Calvin cycle, the pentose phosphate pathway, phosphate acquisition, and nucleotide biosynthesis. (b) Infected *Prochlorococcus* metabolism may be augmented for producing DNA for phage replication. Cyanophage carry AMGs for four of the core pathways. Notably, they appear to lack genes for the Calvin cycle and instead carry an inhibitor of this pathway. The four remaining pathways may be sufficient, however, to fuel DNA production for phage DNA replication, with the products of the light reactions feeding directly into nucleotide biosynthesis.

dicating possibly important functional differences. In Chapter 2, we describe investigations into host and phage transaldolases. Kinetic parameters of phage TalC and host TalB are compared in order to determine if the phage transaldolase has kinetic properties that may provide advantages to an infecting phage.

With information about individual AMGs, we can begin to ask broader questions about their possible roles. *What are the most likely metabolic bottlenecks during infection based on the particular AMGs observed in cyanophage genomes? How does phage infection alter host metabolism, and how is the infection process dependent on the metabolic state of the host?* In Chapter 3, we describe prevalence patterns of AMGs in cyanophage genomes and environmental sequence databases, focusing on genes for the PPP. Gene expression over infection is combined with these prevalence data to derive a model for the role of PPP genes and other AMGs during infection. In Chapter 4, we describe initial efforts to test this model. Phage replication and metabolite level dynamics are measured for infections in the light and in the dark, illuminating the interactions among host, phage, and environment. Finally, in Chapter 5, we describe outstanding questions left by this research and possible experimental approaches to address these questions.

As a final point, we note that this work has leveraged the extensive culture collection of the Chisholm Laboratory in order to use phage and host strains most appropriate for each experiment. Investigations of transaldolase kinetics (Chapter 2) targeted enzymes from a range of phage (podovirus and myovirus) and host (high-light and low-light adapted *Prochlorococcus*) strains, with phage strains and the host strains on which they were isolation studied as pairs. The full library of marine cyanophages infecting *Prochlorococcus* and *Synechococcus* was catalogued for the presence of carbon metabolism genes and other AMGs (Chapter 3). Investigations of phage gene expression (Chapter 3) focused on a cyanophage (Syn9) that carries all four of the PPP-related AMGs (*talC*, *cp12*, *zwf*, and *gnd*) in order to show that the full complement of PPP AMGs is expressed; the optimal host of this phage is a marine *Synechococcus* strain (WH8109) and was therefore used for these experiments. Finally, investigations of metabolite levels during infection (Chapter 4) focused on a *Prochlorococcus* strain (MED4) that has been studied extensively at the transcriptome and proteome level; a phage strain (P-HM2) infecting this host was chosen that carries the two most prevalent PPP AMGs (*talC* and *cp12*) and which yields a robust infection. We would argue that while the use of different strains for different experiments in some ways

complicates a higher-order synthesis, the prevailing features of *Prochlorococcus* and *Synechococcus* metabolism under cyanophage infection are likely conserved, and the availability of multiple strains to address a diverse set of questions should be seen as an asset.

CHAPTER 2

Kinetic and structural properties of cyanophage transaldolase relative to its host *Prochlorococcus* transaldolase

Luke R. Thompson, Alexander U. Singer, Libusha Kelly,
JoAnne Stubbe, and Sallie W. Chisholm
(manuscript to be submitted)

Abstract

Many cyanophages that infect the marine cyanobacteria *Prochlorococcus* and *Synechococcus* carry a gene for transaldolase, a rate-limiting enzyme in the non-oxidative portion of the pentose phosphate pathway. Transaldolase genes have not been documented in viruses other than cyanophages. Phage transaldolase (*talC* gene, TalC protein) and host transaldolase (*talB* gene, TalB protein) have pairwise amino-acid identities ranging from 24–29%, and each has the 14 active-site residues that are universally conserved. However, TalC from cyanophages (∼215 aa) is significantly shorter than TalB from *Prochlorococcus* (∼330 aa), and a structure-based alignment of these sequences shows multiple deletions in TalC relative to TalB. The presence of transaldolase genes in cyanophage led us to question whether these genes encode functional enzymes, whose activity could potentially assist cyanophage reproduction. Further, the differences between phage and host transaldolase sequences led us to question whether they might have significant kinetic differences and whether this could help account for phage encoding an enzyme unlike the host enzyme. To address these questions, we cloned, expressed, and purified transaldolase orthologs from several cyanophages and their host *Prochlorococcus* strains. Kinetic properties of the enzymes were measured spectrophotometrically using a coupled assay. Both the phage and host enzymes

have transaldolase specificity, as predicted by their active-site residues. Turnover numbers of phage TalC were 3.6–5.9 s^{-1} and of *Prochlorococcus* TalB were 14.9–20.8 s^{-1}. Michaelis constants of fructose 6-phosphate (F6P) and erythrose 4-phosphate (E4P) for phage TalC were 0.7–1.6 mM (F6P) and 0.08–0.20 mM (E4P) and for *Prochlorococcus* TalB were 1.0–1.5 mM (F6P) and 0.10–0.15 mM (E4P). A three-dimensional structure of *Prochlorococcus* MIT9312 TalB was determined by x-ray crystallography, and a homology model was made of cyanophage P-SSP7 TalC. Both TalC and TalB structures had a conserved α/β-barrel structure and active-site structure, with the Schiff base-forming lysine (Lys-135, MIT9312 TalB numbering), proton-donating/accepting glutamate (Glu-99) and aspartate (Asp-17), and specificity-determining phenylalanine (Phe-181) in similar positions. The native sizes of these proteins were assessed using size-exclusion chromatography (SEC). SEC indicated that TalC formed a pentamer in solution, whereas TalB formed a monomer. The lower activity of cyanophage TalC relative to *Prochlorococcus* TalB suggests that kinetics are insufficient to explain cyanophage use of a functional yet less-efficient enzyme. Alternate explanations involving phage genome size limitations and host proteome dynamics are considered.

Introduction

Marine picocyanobacteria of the genera *Prochlorococcus* and *Synechococcus* are the numerically dominant photosynthetic organisms in the ocean (Partensky et al. 1999, Bouman et al. 2006, Johnson et al. 2006), contributing a significant portion of primary productivity in the nutrient-poor open ocean (Li et al. 1983, Vaulot et al. 1995). Cyanophage, viruses that infect cyanobacteria, frequently co-occur with *Prochlorococcus* and *Synechococcus* (Waterbury and Valois 1993, Suttle and Chan 1994, DeLong et al. 2006) and have been isolated on cultured *Prochlorococcus* and *Synechococcus* strains (Waterbury and Valois 1993, Sullivan et al. 2003). Over 20 genomes of cyanophages have now been sequenced (Chen and Lu 2002, Sullivan et al. 2005, Mann et al. 2005, Pope et al. 2007, Weigele et al. 2007, Millard et al. 2009, Sullivan et al. ress). Cyanophage genomes are notable for the presence of genes encoding host-like metabolic proteins. We have termed these genes 'auxiliary metabolic genes' because they are thought to support host metabolism during infection to yield more phage progeny (Breitbart et al. 2007, Appendix E). The most well-known phage metabolic genes encode photosynthesis proteins, including the core photosystem II proteins D1 (*psbA*)

and D2 (*psbD*) (Mann et al. 2003, Millard et al. 2004, Lindell et al. 2004, Sullivan et al. 2006). Other metabolic proteins encoded by cyanophages include enzymes involved in the pentose phosphate pathway and DNA biosynthesis (Sullivan et al. 2005, Weigele et al. 2007). The genes for these proteins are transcribed simultaneously and thus are thought to work together during cyanophage infection (Thompson et al., Chapter 3), possibly to aid phage genome replication.

Many cyanophages infecting *Prochlorococcus* and *Synechococcus* carry a gene for the pentose phosphate pathway enzyme transaldolase. This gene, *talC*, is found in 80% of *Prochlorococcus* and *Synechococcus* cyanophage genomes sequenced to date (Thompson et al., Chapter 3). Further, among all viral genomes in GenBank (>3500), transaldolase is found only in cyanophages. Thus, transaldolase appears to be a special adaptation of these viruses for infecting a cyanobacterial host.

Transaldolase (EC 2.2.1.2) reversibly transfers a three-carbon dihydroxyacetone moiety from sedoheptulose 7-phosphate (S7P) to glyceraldehyde 3-phosphate (GAP), generating erythrose 4-phosphate (E4P) and fructose 6-phosphate (F6P) (Equation 2.1). Transaldolase functions in the non-oxidative portion of the pentose phosphate pathway (PPP) (Horecker et al. 1961), in which it is thought to be the rate-limiting step (Heinrich et al. 1976, Banki et al. 1996). As shown in Figure 2-1, transaldolase helps regenerate F6P from ribulose 5-phosphate (Ru5P), allowing the oxidative portion of the PPP to continue oxidizing F6P to Ru5P, carbon dioxide, and NADPH (Berg et al. 2007). The PPP produces carbon skeletons and reducing equivalents for a variety of cellular processes. For example, ribose 5-phosphate (R5P), generated from Ru5P, is an important precursor to ribnonucleotides, which can then be converted to deoxyribonucleotides with NADPH as the reductant (Wood 1986b). For a review of the PPP and its discovery, see Horecker (2002).

$$\text{S7P} + \text{GAP} \rightleftharpoons \text{E4P} + \text{F6P} \tag{2.1}$$

Structurally, transaldolase is a member of the class I aldolase superfamily, the members of which all utilize a Schiff-base intermediate and share the same α/β-barrel (TIM-barrel) structure (Choi et al. 2006). Among the transaldolase family, five phylogenetic subfamilies have been defined (Samland and Sprenger 2009). Notable subfamilies include TalB (subfamily 1) and TalC (subfamily 4), whose members have been shown to have transaldolase

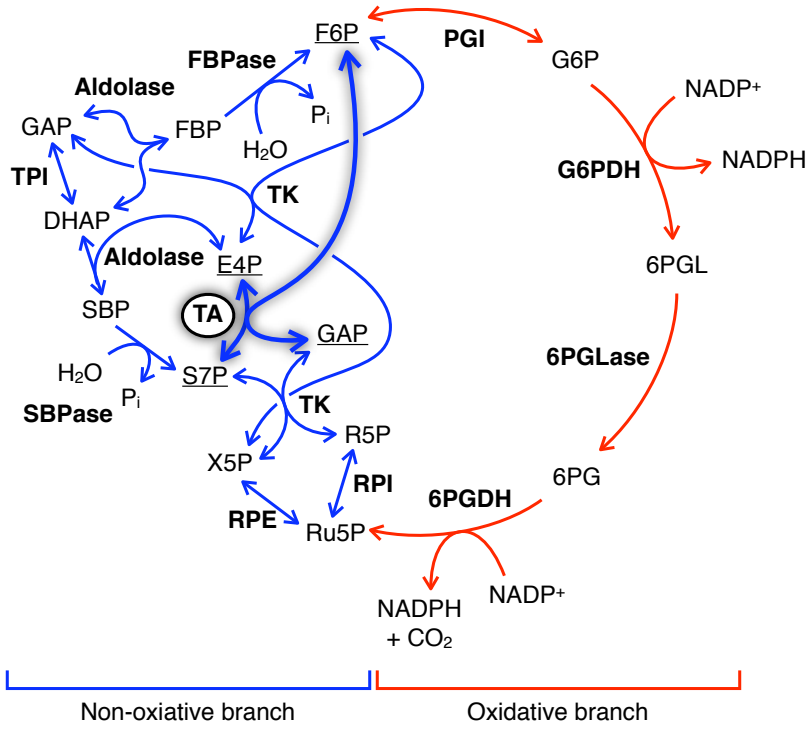


Figure 2-1: Diagram of the pentose phosphate pathway (PPP), showing which enzymes are part of the oxidative portion (red) or the non-oxidative portion (blue). The transaldolase reaction is shown in bold with its substrates underlined. Enzyme abbreviations: aldolase, fructose-1,6-bisphosphate aldolase/sedoheptulose-1,7-bisphosphate aldolase; FBPase, fructose-1,6-bisphosphatase; G6PDH, glucose-6-phosphate dehydrogenase; PGI, phosphoglucose isomerase; PGLase, 6-phosphogluconolactonase; RPE, ribulose-5-phosphate epimerase; RPI, ribulose-5-phosphate isomerase; SBPase, sedoheptulose-1,7-bisphosphatase; 6PGDH, 6-phosphogluconate dehydrogenase; TA, transaldolase; TK, transketolase; TPI, triosephosphate isomerase. Metabolite abbreviations: DHAP, dihydroxyacetone phosphate; E4P, erythrose 4-phosphate; FBP, fructose 1,6-bisphosphate; F6P, fructose 6-phosphate; GAP, glyceraldehyde 3-phosphate; G6P, glucose 6-phosphate; R5P, ribose 5-phosphate; Ru5P, ribulose 5-phosphate; SBP, sedoheptulose 1,7-bisphosphate; 6PG, 6-phosphogluconate; 6PGL, 6-phosphogluconolactone; S7P, sedoheptulose 7-phosphate; X5P, xylulose 5-phosphate.

activity (Cremona et al. 1965, Sprenger et al. 1995, Schürmann and Sprenger 2001, Soderberg and Alver 2004), and Fsa (subfamily 5), whose members have been shown to have fructose-6-phosphate aldolase activity (Schürmann and Sprenger 2001). Despite different activities, TalC and Fsa types are structurally more similar to each other than to TalB. TalB and TalC/Fsa are easily distinguished by primary structure: they share only 22–36% sequence identity, and the 'classical' TalB transaldolases (∼310–350 aa) are much longer than the TalC transaldolases or Fsa fructose-6-phosphate aldolases (∼210–230 aa) (Samland and Sprenger 2009). Deletions in TalC and Fsa correspond to several missing α helices and loops relative to TalB (Thorell et al. 2002). Oligomerization states are also different: TalB subunits usually form monomers or dimers in solution, while TalC or Fsa subunits tend to form decamers (Samland and Sprenger 2009).

Interestingly, transaldolases from *Prochlorococcus* and cyanophages are not from the same subfamilies: *Prochlorococcus* transaldolase is a TalB type, whereas cyanophage transaldolase is a TalC type. Thus, while *Prochlorococcus* and cyanophage transaldolases likely share the Schiff-base mechanism and α/β-barrel fold of the class I aldolase superfamily, as well as the transaldolase specificity found in the TalB and TalC subfamilies, they share less than 30% amino-acid identity and are predicted to have significant differences in tertiary and quaternary structure. Given the differences between phage and host transaldolases, we asked the following: Do phage and host transaldolases have different kinetic properties, and if so, can these kinetic differences explain why phages encode an enzyme distinct from the host enzyme? To address this question, we cloned, expressed, and purified transaldolases from three *Prochlorococcus* strains and three cyanophages known to infect those strains. We measured activities in vitro using a coupled spectrophotometric assay to determine their kinetic properties. We also used x-ray crystallography, homology modeling, and size-exclusion chromatography (SEC) to investigate the structural differences between phage and host transaldolases.

Materials & Methods

Materials

DNeasy Blood and Tissue kits and Miniprep Spin kits were from QIAGEN (Valencia, CA, USA). Phusion High-Fidelity Polymerase and dNTPs were from New England Biolabs (Ip-

swich, MA, USA). Oligonucleotides were from Integrated DNA Technologies (Coralville, IA, USA). Champion pET Directional TOPO vectors and *E. coli* BL21 Star (DE3) One Shot chemically competent cells were from Invitrogen (Carlsbad, CA, USA). BL21-CodonPlus (DE3)-RIPL competent cells, ArcticExpress (DE3) competent cells, and QuikChange Site-Directed Mutagenesis kits were from Stratagene (La Jolla, CA, USA). LB medium was from Becton, Dickinson and Company (Franklin Lakes, NJ, USA). Ampicillin, chloramphenicol, gentamicin, glycylglycine (Gly-Gly), imidazole, hen egg lysozyme, Sephadex G-25 medium resin, sodium dodecyl sulfate (SDS), Bradford reagent, bovine serum albumin (BSA), fructose 6-phosphate (∼98% by enzymatic assay, lot no. 015K7013), erythrose 4-phosphate (∼60% by enzymatic assay, lot no. 115K3789), rabbit-muscle triosephosphate isomerase (4400 U/mg, lot no. 035K7457), and rabbit-muscle glycerol-3-phosphate dehydrogenase (270 U/mg, lot no. 035K7457) were from Sigma-Aldrich (St. Louis, MO, USA). Isopropyl β-D-1-thiogalactopyranoside (IPTG) and 1,4-dithiothreitol (DTT) were from Promega (Madison, WI, USA). DNase and Complete Mini EDTA-free protease inhibitor tablets were from Roche (Indianapolis, IN, USA). Nickel-nitrilotriacetic acid (Ni-NTA) agarose resin was from QIAGEN (Valencia, CA, USA). Amicon Ultra-15 centrifugal filter units were from Millipore (Billerica, MA, USA). Q Sepharose Fast Flow anion-exchange resin was from Amersham (Piscataway, NJ, USA). Slide-A-Lyzer dialysis cassettes were from Pierce (Rockford, IL, USA). BugBuster protein extraction reagent was from Novagen (Darmstadt, Germany). Acrylamide-bisacrylamide, Laemmli buffer, β-mercaptoethanol, Mini PROTEAN 3 gel apparatus, and gel filtration molecular weight standards were from Bio-Rad (Hercules, CA, USA). Ammonium persulfate (APS) was from Mallinckrodt Baker (Phillipsburg, NJ, USA). Reduced β-nicotinamide adenine dinucleotide (NADH) was from Calbiochem (Gibbstown, NJ, USA). Paratone-N oil was from Hampton Research (Aliso Viejo, CA, USA).

Cloning of recombinant transaldolases

Prochlorococcus and cyanophage genomic DNA was isolated using a DNeasy Blood and Tissue kit (cells) or used directly (phage) and amplified by PCR (polymerase chain reaction) using Phusion High-fidelity polymerase with primer sequences given in Table 2.1. Amplicons were cloned into Champion pET Directional TOPO vector pET100 or pET101 or plasmid p15TvLic. Plasmid p15TvLic is a modified version of vector pET-15b from Novagen (Darmstadt, Germany) in which the TEV protease cleavage site replaces the thrombin cleavage site

Table 2.1: PCR primers used in this study. MIT9312, MED4, and NATL2A are *Prochlorococcus* strains; P-SSM2, P-SSM4, and P-SSP7 are cyanophage strains.

Gene	Vector	F/R	Primer sequence
MIT9312 *talB*	p15TvLic	F	5'-TTGTATTTCCAGGGCATGAAATCAATTTTAGAACAATTGTC-3'
		R	5'-CAAGCTTCGTCATCAGTTGGCAGAAATTAATTTATGATTTTTCA-3'
MED4 *talB*	pET100	F	5'-CACCATGAAATCAATTTTAGAACAATTATC-3'
		R	5'-CTAAGTTGTCGAAATTAATTTTTGATTATTAATTTC-3'
NATL2A *talB*	pET100	F	5'-CACCATGGAATCCCTGCTGAGTCAGCTGTC-3'
		R	5'-TCAGTGAGTTAGGGCAACTTCTCC-3'
P-SSM2 *talC*	pET101	F	5'-CACCATGAAAATCTTTTTAGATACTGCC-3'
		R	5'-ACGCTTAACCTGAGCCC-3'
P-SSM4 *talC*	pET101	F	5'-CACCATGAAACTATTTTTAGATTGTTCAG-3'
		R	5'-TCCTCCTACAAGTTTAGTCCAATC-3'
P-SSP7 *talC*	pET101	F	5'-CACCATGAAGATATTTCTGGATTCAG-3'
		R	5'-GACATTTCTGCCAAAATCTAAGGC-3'

Table 2.2: Cloning vectors for the recombinant transaldolases. MIT9312, MED4, and NATL2A are *Prochlorococcus* strains; P-SSM2, P-SSM4, and P-SSP7 are cyanophage strains. The TEV cleavage site of MIT9312 TalB is marked with a down arrow.

Vector	Source	His-tag	Protein	Size (kDa)
p15TvLic	Novagen/ Zhang et al. (2001)	N-terminal 6×His MGSSHHHHHHSSGENLYFQ↓G...	MIT9312 TalB	39.7
pET100	Invitrogen	N-terminal 6×His MRGSHHHHHHGMASMTGGQQMGRDLYDDDDKDHPFT...	MED4 TalB NATL2A TalB	41.3 40.6
pET101	Invitrogen	C-terminal 6×His ...KGELNSKLEGKPIPNPLLGLDSTRTGHHHHHH	P-SSM2 TalC P-SSM4 TalC P-SSP7 TalC	27.4 27.1 27.2

and a double stop codon is inserted downstream of the BamHI site (Zhang et al. 2001). Additional information on constructs is contained in Table 2.2. Site-directed mutagenesis was performed using the QuikChange site-directed mutagenesis kit. Sequences of *talB* and *talC* constructs were confirmed by DNA sequencing at the Massachusetts Institute of Technology Biopolymers Laboratory. Cyanophage *talC* pET101 constructs were transformed into *E. coli* BL21 Star (DE3) One Shot chemically competent cells, *Prochlorococcus* MIT9312 *talB* p15TvLic construct was transformed into BL21-CodonPlus (DE3)-RIPL competent cells, and *Prochlorococcus talB* pET100 constructs were transformed into ArcticExpress (DE3) competent cells.

Expression and purification of *Prochlorococcus* TalB

The expression and purification of *Prochlorococcus* MED4 TalB is described here. Nearly identical expression and purification protocols were used for *Prochlorococcus* MIT9312 TalB and *Prochlorococcus* NATL2A TalB. Cells carrying the pET constructs were grown up in LB medium containing 100 μg/mL ampicillin and 20 μg/mL gentamicin with shaking at 37°C. When OD_{600} reached 0.5, temperature was reduced to 13°C and cultures incubated until OD_{600} was 1.0 (∼3 h), whereupon they were induced with 0.5 mM IPTG. Following 35 h of growth at 13°C, cells were harvested by centrifugation for 10 min at 3,000×g. Typical yield was 4 g cell paste per liter of culture.

Cell paste (16 g) was resuspended in 80 mL of buffer A (50 mM Gly-Gly (pH 8.0), 500 mM NaCl, and 5% glycerol) with 10 mM imidazole, 1 mg/mL lysozyme, 10 units/mL DNase, and 2 Complete Mini EDTA-free protease inhibitor tablets. This resuspension was lysed using a FRENCH pressure cell press (Thermo Scientific, Waltham, MA, USA) at 14,000 psi, and the lysate was centrifuged for 10 min at 40,000×g. The supernatant fraction was incubated with Ni-NTA resin (16 mL) and buffer A with 10 mM imidazole in a final volume of 160 mL. The slurry was poured into a column and material unbound to the resin allowed to flow through, followed by washes of 40 column volumes (CV) of buffer A containing 10 mM imidazole. Bound proteins were eluted with a 100×100-mL linear gradient of 10–300 mM imidazole in Buffer A. TalB eluted at 120 mM imidazole, and fractions with high protein content (based on A_{280}) or high activity were pooled and concentrated with Amicon Ultra-15 centrifugal filter units. Concentrated protein was diluted 1:100 to reduce the salt concentration and this solution was loaded onto a Q Sepharose Fast Flow anion-exchange column (5.5×6.5 cm, 150 mL) preequilibrated with 50 mM Gly-Gly (pH 8.0) containing 5 mM NaCl. Protein was eluted using a 250×250-mL linear gradient of 5–500 mM NaCl in 50 mM Gly-Gly (pH 8.0). TalB eluted at 250 mM NaCl, and fractions with high A_{280} or high activity were pooled and concentrated. The concentrated protein was loaded onto a Sephadex G-25 size-exclusion column (2.5×41 cm, 200 mL) preequilibrated with 50 mM Gly-Gly (pH 8.0). Fractions with high A_{280} were pooled and concentrated. Purified TalB aliquots were stored in 10% glycerol at –80°C.

Expression and purification of cyanophage TalC

The expression and purification of cyanophage P-SSM2 TalC is described here. Nearly identical expression and purification protocols were used for cyanophage P-SSM4 TalC and cyanophage P-SSP7 TalC. Cells carrying the pET constructs were grown up in LB medium containing 100 μg/mL ampicillin with shaking at 37°C. When OD_{600} reached 0.7, temperature was reduced to 25°C and cultures incubated until OD_{600} was 1.0 ($\sim$1 h), whereupon they were induced with 0.5 mM IPTG. Following 15 h of growth at 25°C, cells were harvested by centrifugation for 10 min at 3,000$\times g$. Typical yield was 5 g cell paste per liter of culture.

Cell paste (20 g) was resuspended in 100 mL of buffer A with 20 mM imidazole, 1 mg/mL lysozyme, 10 units/mL DNase, and 2 Complete Mini EDTA-free protease inhibitor tablets. This resuspension was lysed using a FRENCH pressure cell press at 14,000 psi, and the lysate was centrifuged for 10 min at 40,000$\times g$. The supernatant fraction was incubated with Ni-NTA resin (40 mL) and buffer A with 20 mM imidazole in a final volume of 400 mL. The slurry was poured into a column and material unbound to the resin allowed to flow through, followed by washes of 40 CV of buffer A containing 20 mM imidazole. Bound proteins were eluted with a 200×200-mL linear gradient of 20–500 mM imidazole. TalC eluted at 160 mM imidazole, and fractions with high protein content (based on A_{280}) or high activity were pooled and concentrated with Amicon Ultra-15 centrifugal filter units. The concentrated protein was transferred into 50 mM Gly-Gly (pH 8.0) using dialysis with a Slide-A-Lyzer dialysis cassette. Purified TalC aliquots were stored in 10% glycerol at –80°C.

SDS-PAGE and Bradford assays

Subunit molecular weight and distribution of protein between pellet and supernatant fractions after cell lysis was determined by a standard procedure using sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE). Aliquots of approximately 1 mL dense culture were pelleted and resuspended in 100 μL BugBuster protein extraction reagent containing 50 mM Gly-Gly (pH 8.0), 500 mM NaCl, 1 mg/mL lysozyme, 10 units/mL DNase, and 1 tablet/50 mL Complete Mini EDTA-free protease inhibitor tablets. Lysis was achieved by rocking for 10 min at room temperature. Lysate, pellet, and supernatant

fractions were analyzed by SDS-PAGE (12.5%) using the procedure of Laemmli (1970).

Protein concentrations were determined using the method of Bradford (1976). Bovine serum albumin (BSA) of concentrations ranging from 0–1.4 mg/mL (7 μL) or multiple dilutions of transaldolase proteins of unknown concentration (7 μL) were added to a 96-well microtiter plate. Bradford reagent (200 μL) was added simultaneously to all samples, and the plate was incubated at room temperature for approximately 10 min. A_{630} was measured using an Ultramark Microplate Reader (Bio-Rad). A standard curve of A_{630} versus BSA concentration was used to estimate transaldolase concentrations.

Transaldolase assay

Transaldolase activity was measured using a coupled assay as shown in Figure 2-2 and described previously (Bergmeyer et al. 1974). A typical assay in a final volume of 500 μL contained 50 mM Gly-Gly (pH 8.0), 15 mM $MgCl_2$, 10 mM F6P, 1 mM E4P, 0.2 mM NADH, 10 mM DTT, 0.6 U triosephosphate isomerase (TPI), 0.06 U glycerol-3-phosphate dehydrogenase (G3PDH), and transaldolase (approximately 0.5 μg or 0.005 U). NADH consumption was measured by A_{340} ($\epsilon = 6.2$ mM^{-1} cm^{-1}) with a Cary 3 UV–visible spectrophotometer (Varian, Palo Alto, CA, USA) or an Ultramark Microplate Reader (Bio-Rad). Cuvette path length for Cary 3 assays was 1 cm, whereas path length for microtiter plate wells was determined empirically with NADH standards, and a correction factor of 1.785 was applied to Ultramark A_{340} measurements to make them comparable to Cary 3 A_{340} measurements.

For assays using the Cary 3, a solution containing Gly-Gly, $MgCl_2$, F6P, E4P, NADH, and DTT (480 μL, final concentrations as above) was equilibrated at 25°C and monitored for 1 min to confirm no change in A_{340}. A solution of TPI and G3PDH (10 μL, final concentrations as above) was added and any change in A_{340} allowed to dissipate, approximately 5 min. Transaldolase (10 μL) was added and the change in A_{340} monitored for 5 min. For assays using the Ultramark, two microtiter plates were used, and the assay volume was reduced to 200 μL. In plate A, buffer, F6P, and transaldolase (180 μL, final concentrations as above) were premixed and incubated at 25°C for 10 min. In plate B, E4P, NADH, TPI, and G3PDH (20μL, final concentrations as above) were incubated at room temperature for 10 min. The contents of the plate A (180 μL) were added to plate B to initiate the assay.

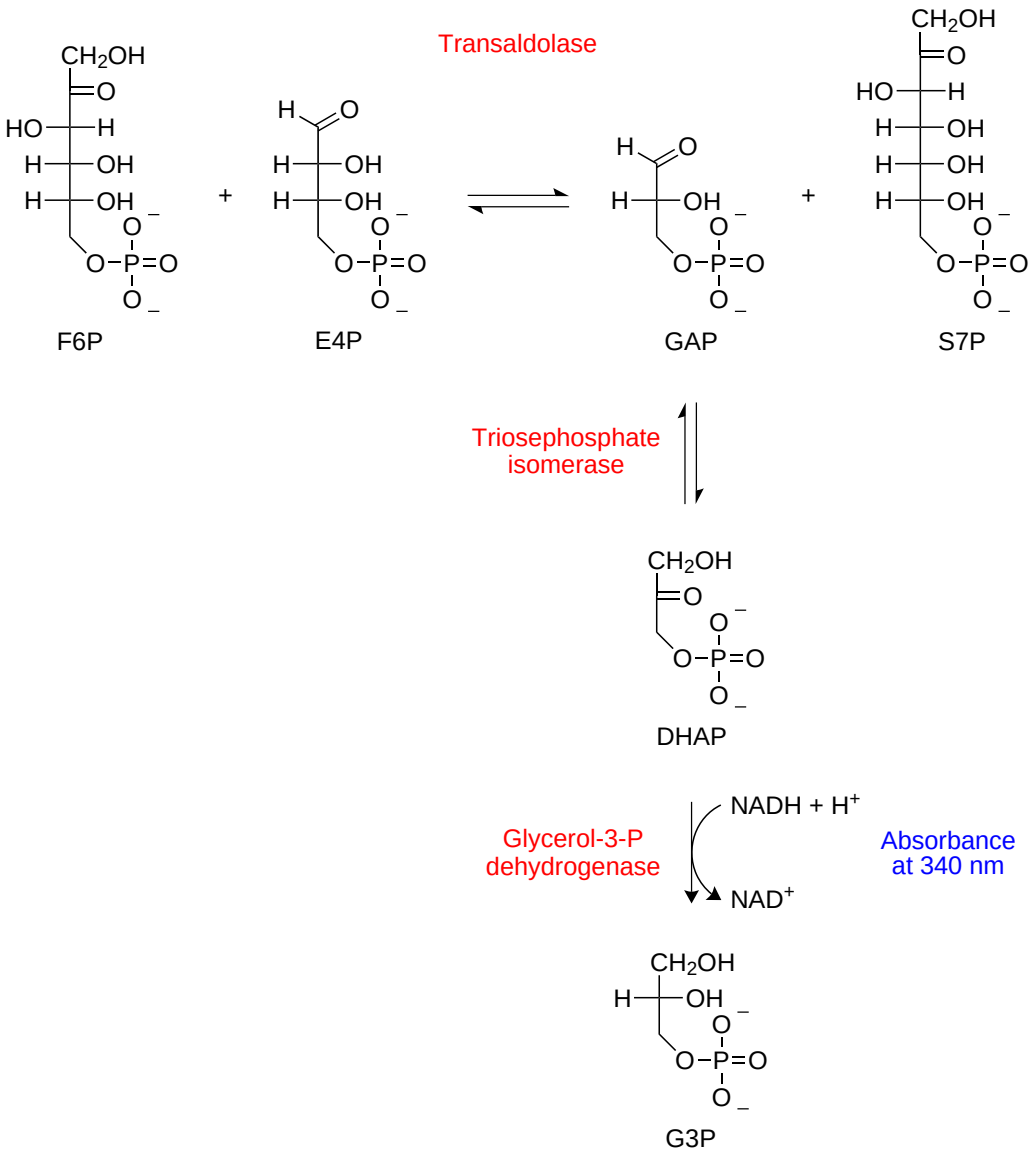


Figure 2-2: Reaction sequence of the transaldolase assay. Transaldolase transfers a dihydroxyacetone moiety from F6P (carbons 1–3) to E4P, generating S7P and GAP. Note that this reaction is reversible, and in nature the substrates are often S7P and GAP and the products F6P and E4P. In the transaldolase assay, GAP production is detected via conversion to DHAP and then G3P, which consumes NADH, and the decrease in NADH is detected spectrophotometrically at 340 nm.

Endpoint assay

E4P and F6P as purchased were 60% and 98% pure, respectively. Actual concentrations for the kinetic analysis described below were determined by endpoint assays (Figure 2-2). Endpoint assays of E4P were carried out with 10 mM F6P and approximately 100 μM E4P; endpoint assays of F6P were carried out with 1 mM E4P and approximately 100 μM F6P. Following dissipation of NADH consumption associated with contaminating GAP (explained below), transaldolase was added and A_{340} monitored. After 20 min (F6P endpoint) or 60 min (E4P endpoint) of incubation, the change in A_{340} was nearly zero, signifying consumption of most of the F6P or E4P. The curves of A_{340} versus time (GAP background curve and F6P or E4P endpoint curve) were fitted to Equation 2.2, a general exponential-decay equation where a, b, and c are constants and t is time in minutes, using linear least-squares analysis implemented with MATLAB software (MathWorks, Natick, MA, USA). This allowed determination of the original and final A_{340}, and the difference between these two values was used to calculate the concentration of NADH consumed and therefore the concentration of the limiting substrate.

$$A_{340} = ae^{-bt} + c \tag{2.2}$$

Determination of kinetic parameters

Kinetic parameters were determined using Equation 2.3, where F6P (10 mM) or E4P (1 mM) was kept constant and the other substrate was varied from 0.05–20 $\times$ K_{m}. Kinetic data were fitted to Equation 2.3 using linear least-squares analysis implemented with MATLAB software.

$$v = \frac{V_{\max}[\mathrm{S}]}{K_{\mathrm{m}} + [\mathrm{S}]} \tag{2.3}$$

Crystallization conditions

Prochlorococcus MIT9312 TalB (19 mg/mL) in 0.2 M HEPES (pH 7.5), 500 mM NaCl, and 0.5 mM tris(2-carboxyethyl)phosphine (TCEP) was transferred to 10 mM HEPES (pH 7.5), 300 mM NaCl, and 0.5 mM TCEP by dialysis and concentrated to 19 mg/mL. Crystallization was performed at room temperature (21°C), to which a home-made preparation

of tobacco etch virus (TEV) protease was added in a ratio of protease to TalB of 1:20 to remove the His-tag from TalB; the protease was not removed prior to crystallization. Crystallization trials were performed using hanging-drop vapor diffusion with an optimized sparse matrix crystallization screen (Kimber et al. 2003). The crystal used for the data collection (see Table 2.4) was obtained using a crystallization liquor containing final concentrations of 20% PEG10K and 0.1 M HEPES (pH 7.5). Crystals were cryoprotected in 20% PEG10K, 0.1 M HEPES (pH 7.5), and 10% ethylene glycol, then rinsed in Paratone-N oil and immediately flash-frozen in liquid nitrogen and stored in liquid nitrogen prior to data collection.

Data collection, structure determination, and refinement

Diffraction data were collected at 100°K on a Rigaku Micromax-007 rotating anode generator equipped with Osmic mirrors. Diffraction data were recorded on an R-Axis IV++ detector and integrated and scaled using HKL2000 (Minor et al. 2006). The structure of *Prochlorococcus* MIT9312 TalB was solved by molecular replacement using the coordinates of human TALDO1 (PDB accession code 1F05) (Thorell et al. 2000) as the initial model. The program PHASER (McCoy et al. 2005), as part of the CCP4 program suite (Collaborative Computational Project 1994), was used to find the position of TalB monomer in the unit cell. The model was then improved by alternate cycles of manual building and water-picking using COOT (Emsley and Cowtan 2004) and restrained refinement against a maximum-likelihood target with 5% of the reflections randomly excluded as an R_{free} test set. All refinement steps were performed using REFMAC (Murshudov et al. 1997) in the CCP4 program suite. Only one residue (residue 333) of the 333-residue protein was omitted in the model due to poor electron density. The final model contains one molecule of TalB, 341 water molecules, and 45 atoms from other small molecules and ions, and was refined to an R_{work} and R_{free} of 15.6% and 20.3%, respectively. Data collection, phasing, and structure refinement statistics are summarized in Table 2.4. The Ramachandran plot generated by PROCHECK (Laskowski et al. 1993) showed excellent stereochemistry overall with 100% of the residues in the most favored and additional allowed regions. The atomic coordinates and structure factors for *Prochlorococcus* MIT9312 TalB have been deposited in the RCSB Protein Data Bank (PDB accession code 3HJZ).

Sequence alignment and phylogenetics

Transaldolase protein sequences were downloaded from GenBank (Benson et al. 2008). The structure-based multiple sequence alignment built by Thorell et al. (2002) was used as a profile alignment in ClustalW 1.83 (Thompson et al. 1994), and *Prochlorococcus* and cyanophage transaldolase sequences were aligned to this profile alignment. For phylogenetic analysis, positions with gaps were removed if they were present in at least 50% of sequences. A tree was built using the maximum likelihood algorithm implemented by PhyML (Guindon and Gascuel 2003), and statistical tests of branches were done using aLRT (approximate likelihood-ratio test) parametric statistics with Chi2-based parametric branch supports (Anisimova and Gascuel 2006). The following parameters were used: phyml_alrt alignment.phy 1 i 1 -2 JTT 0.0 4 e BIONJ y y. The tree was then midpoint rooted and displayed using TreeView (Page 2002).

Structure homology modeling and alignment

Homology models were built using the ClustalW multiple sequence alignment described above and existing crystal structures using the alignment mode of the SWISS-MODEL workspace (Schwede et al. 2003, Arnold et al. 2006). Sequences of TalC from cyanophages P-SSM2, P-SSM4, and P-SSP7 were modeled using the structure of *T. maritima* TalC (Joint Center for Structural Genomics; PDB accession code 1VPX), and sequences of TalB from *Prochlorococcus* MED4 and NATL2A were modeled using the structure of *Prochlorococcus* MIT9312 (this study; PDB accession code 3HJZ).

For visualization of superimposed three-dimensional structures, structure alignments and molecular graphics images were produced using the UCSF Chimera package (Pettersen et al. 2004). Alignments were done using the MatchMaker tool in Chimera, with best-aligning pairs of chains aligned using the Needleman–Wunsch algorithm and the BLOSUM-30 scoring matrix, depending on the average sequence identity among pairs of sequences. Raytraced images were produced with Persistence of Vision Raytracer (v. 3.6) computer software (http://www.povray.org/).

SEC determination of oligomerization state

SEC was performed by using a Superose 12 column (10×300 mm, GE Healthcare, Little Chalfont, UK) attached to a Waters 2487 HPLC. Gel filtration molecular weight standards were vitamin B_{12} (1.35 kDa), myoglobin (17 kDa), ovalbumin (44 kDa), gamma-globulin (158 kDa), and thyroglobulin (670 kDa). The elution buffer was 50 mM Gly-Gly (pH 8.0), 150 mM NaCl. Molecular mass standards were run at the beginning of each experiment. TalB or TalC of concentration 10 mg/mL and volume 40 μL was injected onto the column. The flow rate was 0.5 mL/min. A_{280} was monitored. Plots of log(molecular weight standard) versus retention time were used to estimate molecular weights of TalB and TalC.

Results

Comparative sequence analysis of *Prochlorococcus* and phage transaldolases

Sequence alignment and phylogenetic analysis of *Prochlorococcus* and cyanophage transaldolases reveal key differences that are hallmarks of the TalB and TalC subfamilies (Figure 2-3). The optimized multiple sequence alignment incorporated structural information from *E. coli* FsaA and TalB (Jia et al. 1997, Thorell et al. 2002) and revealed multiple gaps in the phage protein relative to the host protein (Figure 2-3a). Predicted active-site residues are nevertheless conserved between *Prochlorococcus* TalB and cyanophage TalC. Active-site positions in transaldolase, based on multiple structural and kinetic studies (Jia et al. 1997, Schörken et al. 2001, Thorell et al. 2002), are highlighted in Figure 2-3a: absolutely conserved active-site residues are in gray, and variable (though highly conserved) active-site residues are in yellow. *Prochlorococcus* and cyanophage transaldolases are identical at all 14 active-site positions. The only sequence which deviates from the others at imperfectly conserved active-site positions is *E. coli* FsaA. Pairwise sequence identities are detailed in Figure 2-3b: among the three *Prochlorococcus* TalB proteins (blue), sequence identity is 59–86%; among the three cyanophage TalC proteins (red), sequence identity is 45–51%; and between *Prochlorococcus* TalB and cyanophage TalC proteins (purple), sequence identity is 24–29%. *Prochlorococcus* MIT9312 TalB, used as a template for *Prochlorococcus* TalB homology modeling, is 59–86% identical to those proteins; *T. maritima* TalC, used as a

template for cyanophage TalC homology modeling, is 41–44% identical to those proteins. Phylogenetic analysis using the maximum likelihood algorithm (Figure 2-3c) confirms the distinct differences between phage and host transaldolases. *Prochlorococcus* TalB clusters with *E. coli* TalB, and cyanophage TalC clusters with *T. maritima* TalC and *E. coli* FsaA. The results of this comparative sequence analysis thus confirm the significant differences between host and phage transaldolases, supporting the notion that these enzymes may have different kinetic properties and providing motivation for overexpressing and purifying them for kinetic characterization.

Purification of *Prochlorococcus* and phage transaldolases

TalB from *Prochlorococcus* MIT9312, MED4, and NATL2A was purified to 95% homogeneity (Figure 2-4), and TalC from cyanophages P-SSM2, P-SSM4, and P-SSP7 was purified to 90% homogeneity (Figure 2-4), by a procedure outlined in methods. A typical preparation of *Prochlorococcus* TalB yielded 2.5 mg TalB per gram of cell paste with a specific activity of 22 μmol min^{-1} mg^{-1}. A typical preparation of cyanophage TalC yielded 18 mg TalC per gram of cell paste with a specific activity of 10 μmol min^{-1} mg^{-1}.

Overexpression of soluble TalC from cyanophages was readily achieved in BL21 Star (DE3) One Shot cells. Overexpression of soluble TalB from *Prochlorococcus*, however, was more difficult to achieve; at various growth temperatures and IPTG concentrations with BL21 Star (DE3) One Shot cells or BL21-CodonPlus (DE3)-RIPL cells, no soluble protein was isolated. This insolubility of proteins was seen with all three TalBs. A typical example of solubility problems is shown in Figure 2-5a. Lanes 3 and 5 show induced but insoluble *Prochlorococcus* NATL2A TalB in the crude lysate. The supernatant contains no detectable protein of the proper size (∼41 kDa).

To address low solubility of *Prochlorococcus* TalB in *E. coli*, we tried several approaches, including variable IPTG concentrations, lower temperature before induction, and a number of expression cell lines. As an example of this multipronged approach, Figure 2-5 shows expression of *Prochlorococcus* NATL2A TalB under two different conditions: CodonPlus RIPL BL21 (DE3) cells grown at 37°C and ArcticExpress BL21 (DE3) cells grown at 13°C. CodonPlus RIPL cells provide additional tRNAs, which can increase expression if the induced gene uses many codons not frequently used by *E. coli* (Stratagene). ArcticExpress cells express cold-adapted GroEL/GroES chaperonins, which aid folding of proteins and

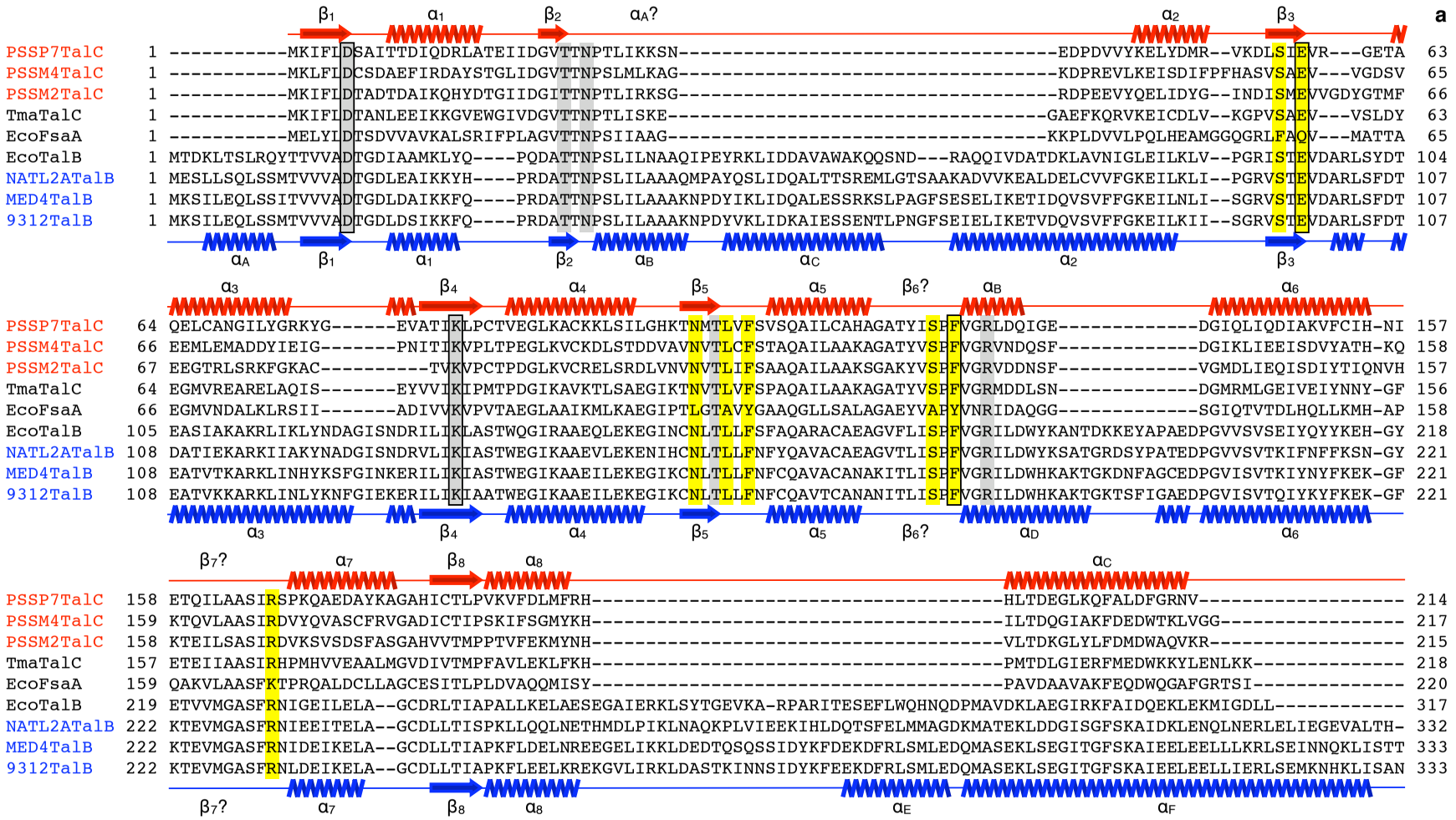


Figure 2-3: Comparison of transaldolase sequences from cyanophages, *Prochlorococcus*, *E. coli*, and *T. maritima*. Names of transaldolases from cyanophages and *Prochlorococcus* are in red and blue, respectively. (a) Multiple sequence alignment, generated starting from the structure-based alignment of Thorell et al. (2002). Active-site residues are colored according to whether they are identical (gray) or variable (yellow), and those pictured in the structure in Figure 2-16b are outlined in black. Secondary structure elements of cyanophage P-SSP7 TalC and *Prochlorococcus* MIT9312 TalB are annotated above and below the alignment and are named as described previously (Thorell et al. 2002). Helix 'α_A?' in P-SSP7 TalC and strands 'β_6?' and 'β_7?' in both structures were expected based on *E. coli* FsaA and *E. coli* TalB but were not predicted in the structure model (MIT9312 TalB) or homology model (P-SSP7 TalC).

b

	TmaTalC	PSSM2TalC	PSSM4TalC	PSSP7TalC	EcoFsaA	EcoTalB	NATL2ATalB	MED4TalB
PSSM2TalC	43							
PSSM4TalC	44	51						
PSSP7TalC	41	45	45					
EcoFsaA	29	25	30	30				
EcoTalB	33	24	29	25	24			
NATL2ATalB	34	24	29	26	22	55		
MED4TalB	36	24	28	26	22	54	62	
9312TalB	36	24	29	26	22	53	59	86

c

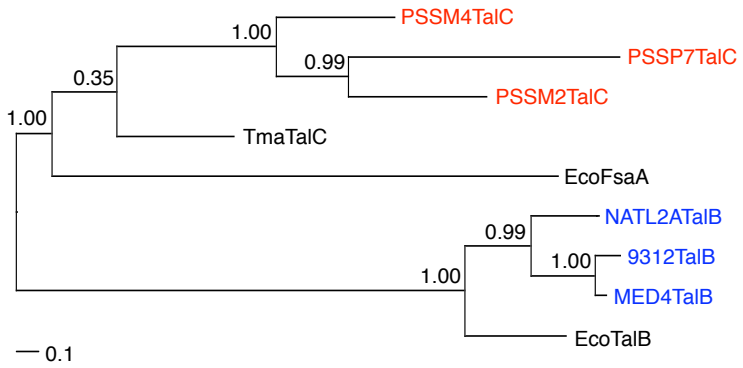


Figure 2-3: (continued) (b) Pairwise amino-acid identities from alignment. Phage pairs are outlined in red, *Prochlorococcus* pairs are outlined in blue, and phage/*Prochlorococcus* pairs are outlined in purple. (c) Maximum likelihood tree generated from alignment, midpoint-rooted with Chi2-based parametric branch supports. Abbreviations: TmaTalC, *T. maritima* TalC; PSSM2TalC, cyanophage P-SSM2 TalC; PSSM4TalC, cyanophage P-SSM4 TalC; PSSP7TalC, cyanophage P-SSP7 TalC; EcoFsaA, *E. coli* FsaA; EcoTalB, *E. coli* TalB; NATL2ATalB, *Prochlorococcus* NATL2A TalB; MED4TalB, *Prochlorococcus* MED4 TalB; 9312TalB, *Prochlorococcus* MIT9312 TalB.

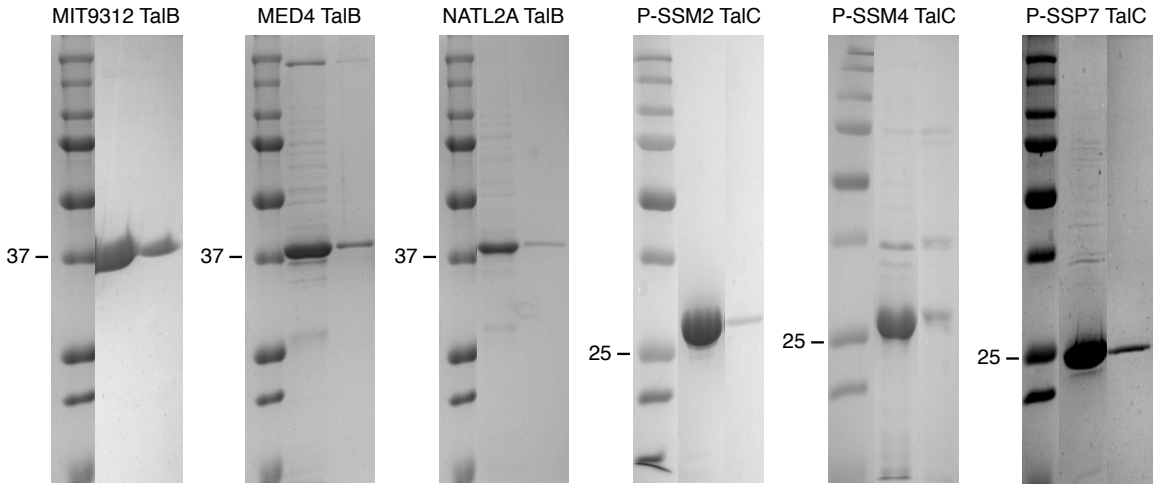


Figure 2-4: SDS-PAGE gels of purified *Prochlorococcus* TalB and cyanophage TalC. Large (10 μg, lane 2) and small (0.5 μg, lane 3) amounts of protein were loaded to determine purity and size, respectively. Molecular weight standards were in each gel (lane 1). Standards of 37 kDa and 25 kDa are labeled.

can refold improperly folded proteins (Stratagene).

As shown in Figure 2-5, adequate solubility of *Prochlorococcus* TalB was achieved with ArcticExpress cells grown at 13°C. Lanes 9 and 10 of Figure 2-5a show expressed NATL2A TalB in both the crude lysate and the supernatant fraction (arrow next to lanes 9–10). This band is the proper size for NATL2A TalB (∼41 kDa), and it is not present in the uninduced sample (t=0, lanes 7–8). We also did a small-scale purification of NATL2A TalB from these same cells. As shown in Figure 2-5b, no soluble protein could be purified with the Ni-NTA affinity column from CodonPlus RIPL cells, in contrast with protein expression from ArcticExpress cells. Protein of the proper size is visible in the wash and fractions 1–4 (arrow next to lanes 8–12).

Optimization of the transaldolase assay

The standard assay for transaldolase is shown in Figure 2-2. A number of problems were encountered and needed to be overcome before the assay was effective. First, in our assays, we noticed that there was always background activity before transaldolase was added. That is, there was significant consumption of NADH, measured by A_{340}, without the addition of transaldolase. We found that this background was observed without addition of F6P but not without addition of E4P. Thus, we attributed the background activity to impurities

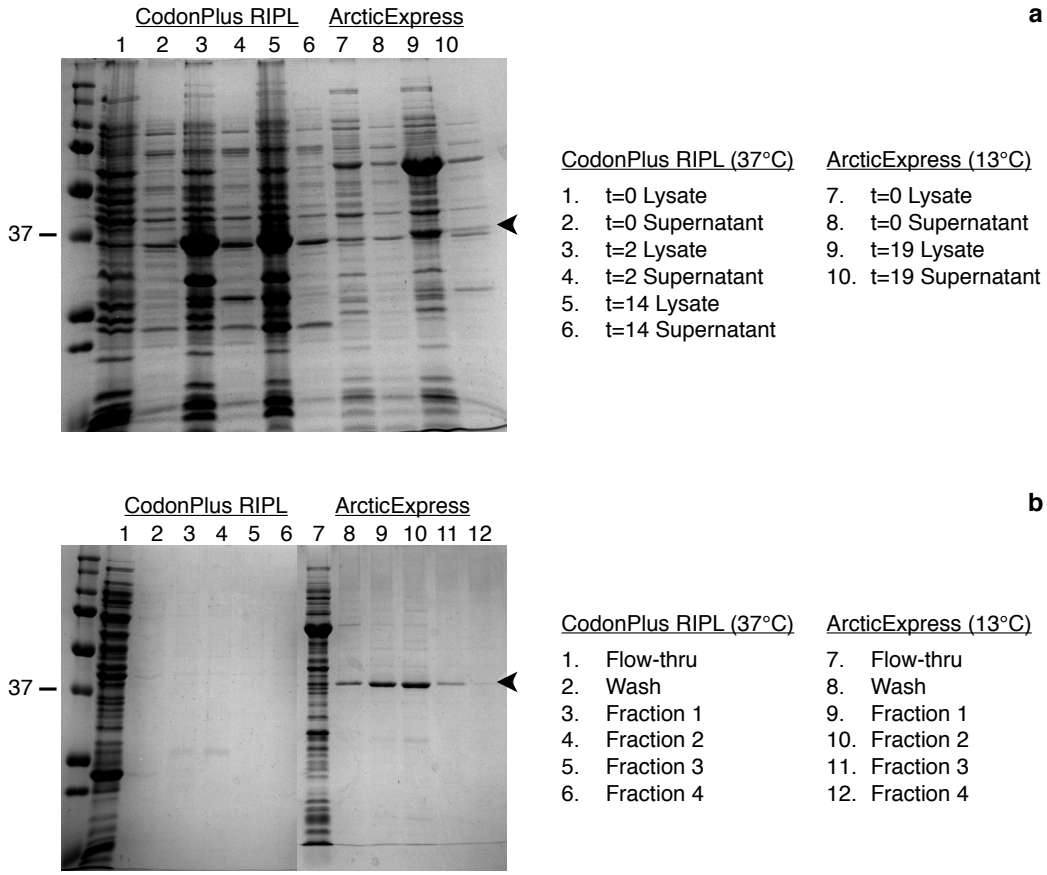


Figure 2-5: SDS-PAGE gels of *Prochlorococcus* NATL2A TalB showing differential solubility in different expression strains and under different growth conditions. (a) Gel showing crude lysate and supernatant fractions of proteins from *E. coli* before (t=0) or after (t=2, t=14, t=19) induction. (b) Purification gel showing flow-through, wash, and elution fractions 1–4. Molecular weight standards were run in the left lane of each gel. The 37-kDa standard is labeled.

in E4P. According to the manufacturer of E4P (Sigma-Aldrich), their preparation of E4P is only around 60% pure. The method they use to produce E4P is that of Ballou (1963), which uses lead tetraacetate oxidation of glucose 6-phosphate and is known to introduce contaminating GAP, around 3% by mass of the final product. Since GAP is measured by the transaldolase assay (Figure 2-2), we therefore attributed the observed background activity to this contaminating GAP. To prevent this contamination from confounding measurements of transaldolase activity, we incubated E4P with TPI and G3PDH until contaminating GAP was converted to G3P (approximately 5 min) and then added transaldolase to start the reaction.

In order to obtain kinetic parameters to describe each enzyme, it was necessary to deter-

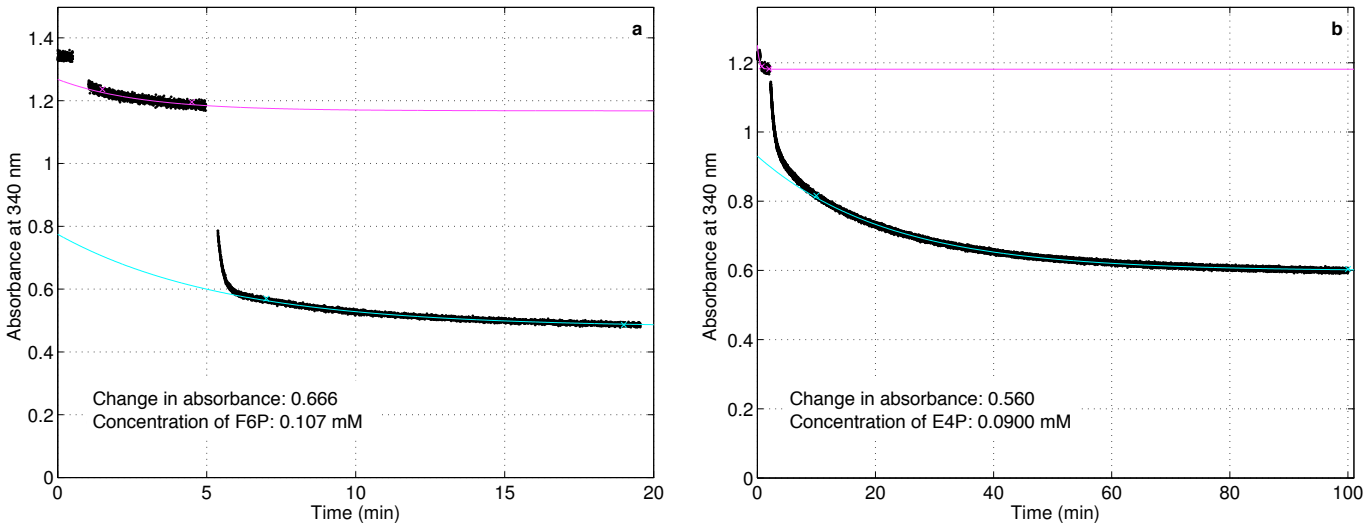


Figure 2-6: Endpoint assays of (a) F6P and (b) E4P. The difference in absorbance between when contaminating glyceraldehyde 3-phosphate is consumed (magenta curve) and when F6P or E4P is consumed (cyan curve) was used to determine the precise concentration of each substrate.

mine the concentrations of F6P and E4P. Due to impurities in E4P and inherent difficulties in measuring masses of small quantities of sugar phosphates, which are hygroscopic, this required an alternative approach. E4P and F6P concentrations were determined enzymatically using endpoint assays (Figure 2-6). As described in the methods, for each endpoint assay, buffer, NADH, and F6P or E4P (one in excess, the other limiting) were added to the cuvette. Coupling enzymes were then added and the change in absorbance measured, corresponding to background activity from contaminating GAP (magenta curve in Figure 2-6). Then transaldolase was added and the change in absorbance monitored, corresponding to total consumption of the limiting substrate (cyan curve in Figure 2-6). Decay curves (magenta and cyan) were fitted to an exponential decay equation to determine the absorbance after total consumption of substrate.

The second problem encountered with the assay in the case of TalB was a long and variable lag phase. *Prochlorococcus* TalB expressed from vector pET100, with N-terminal His-tag MRGSHHHHHHGMASMTGGQQMGRDLYDDDDKDHPFT, showed a lag in rate of turnover before linear/maximal activity could be achieved (Figure 2-7a). When TalB was added last to the transaldolase assay, as was standard in our assays, it regularly took 6–7 minutes for the rate of change in A_{340} to reach its maximum. Through a process of trial and

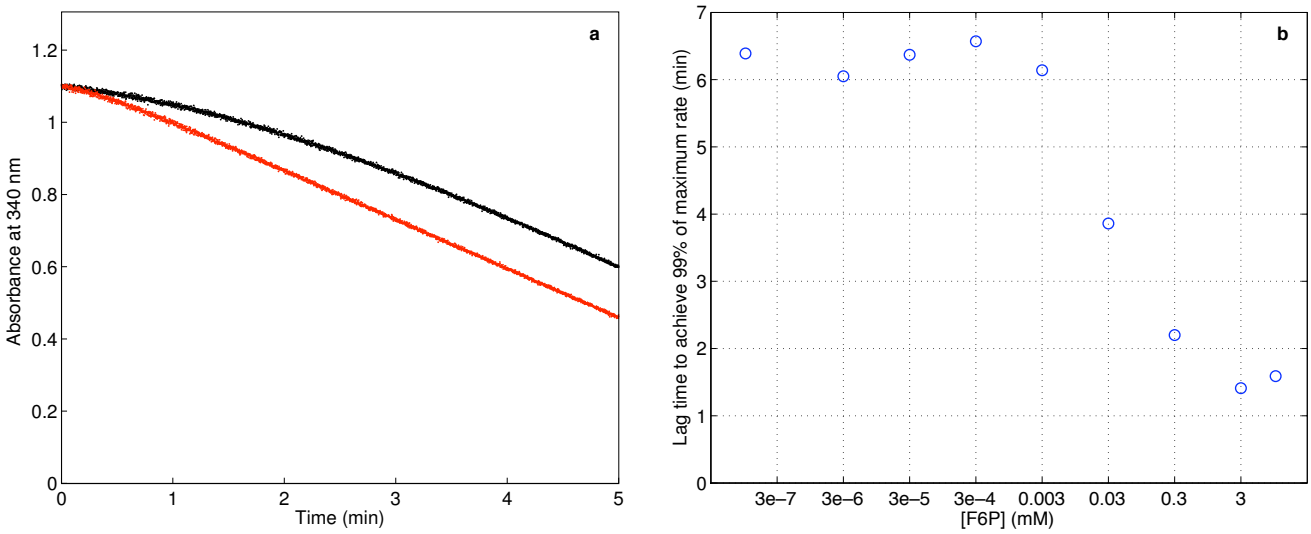


Figure 2-7: Lag in activity of MED4 TalB with pET100 N-terminal His-tag is eliminated by pre-incubation of the enzyme with F6P. Pre-incubations of TalB with F6P were 10 min in duration. (a) Activity of MED4 TalB without any preincubation (black) or following preincubation with 10 mM F6P (red). (b) Preincubation with an F6P concentration in the micromolar–millimolar range is sufficient to reduce the lag in activity. The typical concentration of F6P used for preincubation in assays was 1–10 mM.

error, we determined that the lag phase could be diminished to 1.5 min by incubating TalB with 1 mM F6P for 10 min before initiating the assay. This is shown for *Prochlorococcus* MED4 TalB in Figure 2-7. A similar lag in activity was seen in *Prochlorococcus* NATL2A TalB with the same N-terminal His-tag (vector pET100).

While the requirement of F6P for an optimally active conformation could be interesting, we considered the possibility that the effect was related to the purification tag associated with the expression vector. Using a different vector, p15TvLic, we found the lag in TalB was eliminated. Specifically, *Prochlorococcus* MIT9312 TalB was expressed using the p15TvLic vector. As shown in Table 2.2, the His-tag incorporated by p15TvLic (N-MGSSHHHHHHSSGENLYFQ...) is significantly different from the His-tag incorporated by pET100 (N-MRGSHHHHHHGMASMTGGQQMGRDLYDDDDKDHPFT...). While we did not do a side-by-side comparison of the same ortholog in the two different vectors, we infer that the tag is likely responsible for the unusual kinetic behavior. MIT9312 TalB is 86% identical to MED4 TalB but only 59% identical to NATL2A TalB (Figure 2-3). While pET100 versions of both MED4 TalB and NATL2A TalB proteins exhibit a lag and F6P effect, the p15TvLic version of MIT9312 has no lag or F6P effect. Because sequence simi-

larity is inconsistent with this trend, it appears that the His-tag incorporated by pET100 is responsible for the lag and F6P effect.

Finally, in light of these problems, we wished to confirm that our transaldolase assay was working properly. We tested the assay with *E. coli* TalB that had been cloned with the pET101 construct (C-terminal His-tag as described in Table 2.2), overexpressed, and the enzyme purified to homogeneity by standard procedures. A specific activity of 82 μmol min^{-1} mg^{-1} was obtained that compares favorably with the published specific activity of *E. coli* TalB (60 μmol min^{-1} mg^{-1}) (Sprenger et al. 1995). Using F6P and E4P with concentrations determined by endpoint assay, we obtained K_{m} values of 1.1 mM for F6P and 0.09 mM for E4P. The published K_{m} values are 1.2 mM for F6P and 0.09 mM for E4P (Sprenger et al. 1995). These results for *E. coli* TalB confirmed that the transaldolase assay was working effectively.

Specificity of *Prochlorococcus* and phage transaldolases

Transaldolases, particularly TalC isozymes, are closely related to F6P aldolases (Fsa) (Figure 2-3). As described above, TalC and Fsa share many structural properties, yet their differential activities are associated with differences in several key active-site residues. We predicted based on these differences that the cyanophage enzymes would have transaldolase and not F6P aldolase activity. As shown in Figure 2-8, the mechanisms of transaldolase and F6P aldolase differ in that transaldolase requires E4P as an acceptor substrate for dihydroxyacetone (DHA), whereas F6P aldolase does not. Without E4P, transaldolase cannot turn over and is trapped by the Schiff base of DHA. The transaldolase assay (Figure 2-2) detects GAP production, and without F6P aldolase activity there is no more than stoichiometric GAP production without addition of E4P. In practice, there is some hydrolysis of DHA with transaldolase, but the rate of hydrolysis is much lower than with F6P aldolase. As shown in Figure 2-9 for cyanophage P-SSP7 TalC, there was no turnover without addition of E4P, indicating that this enzyme has transaldolase activity but not F6P aldolase activity. This has been subsequently shown for all cyanophage TalCs and *Prochlorococcus* TalBs.

Effect of DTT on *Prochlorococcus* TalB

Dithiothreitol (DTT) is often included in protein purifications and assays to stabilize proteins by preventing redox chemistry. Initially, DTT was included in all of our transaldol-

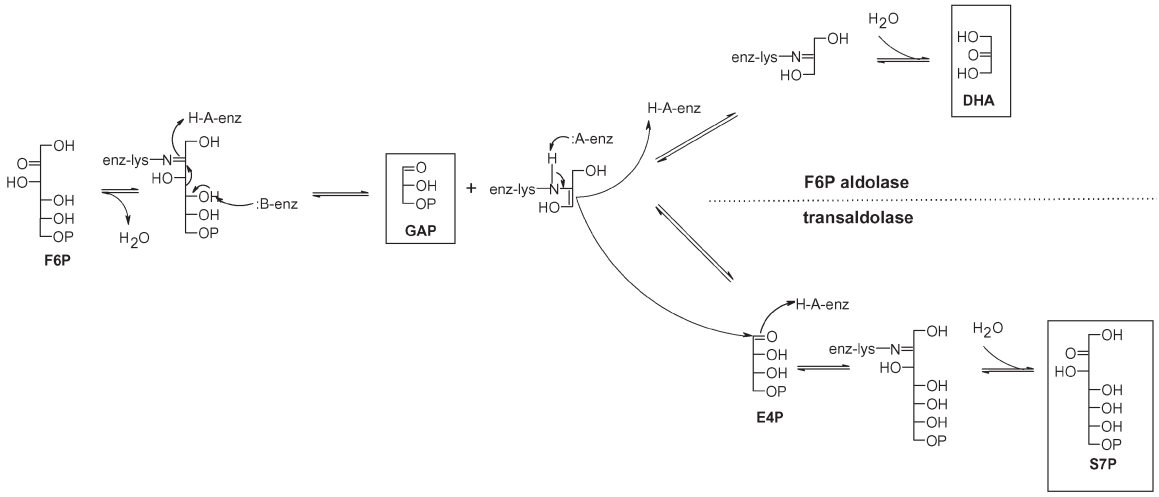


Figure 2-8: Comparison of transaldolase mechanism and F6P aldolase mechanism. Reproduced from Soderberg and Alver (2004).

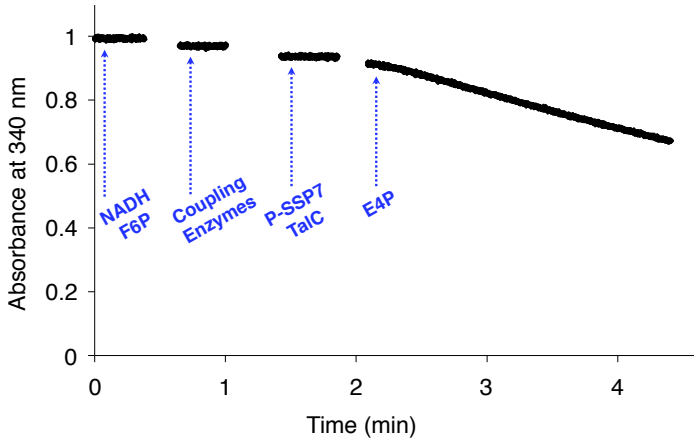


Figure 2-9: Cyanophage P-SSP7 TalC has transaldolase and not F6P aldolase activity. This was shown for all cyanophage and *Prochlorococcus* transaldolases tested. No activity was detected without the addition of E4P, which in transaldolase serves as an acceptor substrate for dihydroxyacetone. Fructose-6-phosphate aldolase requires no acceptor substrate for turnover.

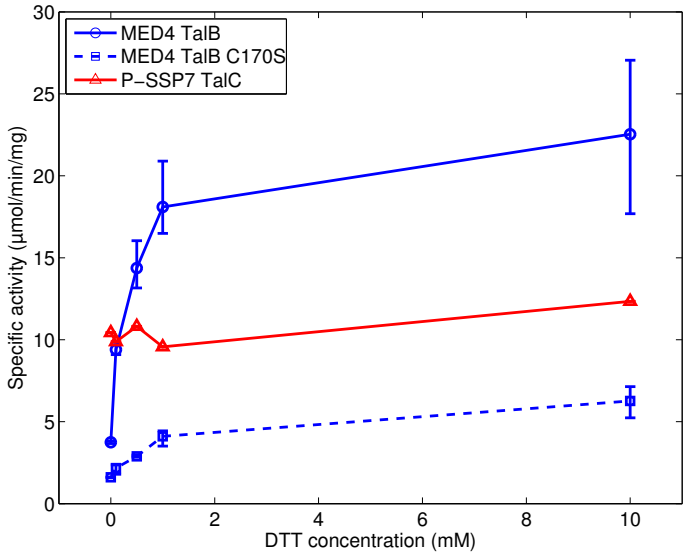


Figure 2-10: Effect of DTT on the activity of TalB expressed from pET100, which incorporates the N-terminal His-tag MRGSHHHHHHGMASMTGGQQMGRDLYDDDDKDHPFT. Error bars represent the maximum and minimum of three replicate assays.

ase assay mixtures. By chance, however, we conducted some assays without DTT in the assay mixture and found, to our surprise, that DTT significantly affected the activity of *Prochlorococcus* TalB. In assays of MED4 TalB and NATL2A TalB, both containing pET100 N-terminal His-tags (Table 2.2), DTT in the range of 1–10 mM was found to increase activity approximately five-fold over assays with no DTT (Figure 2-10). Notably, this effect was not observed in TalCs, which contain pET101 C-terminal His-tags (Table 2.2). When assays were conducted in 10 mM DTT, MED4 TalB activity was about two-fold greater than P-SSP7 TalC, but when assays were conducted in 0 mM DTT, TalC actually had higher activity, around three-fold greater than TalB. This DTT effect appeared to be reversible. When TalB was incubated with excess DTT for 1 h followed by dialysis overnight into buffer with no DTT, activity assays still showed a DTT effect: relative to activity in 0 mM DTT, this 're-oxidized' TalB had activity 50% higher in 1 mM DTT and 90% higher in 10 mM DTT. Purification with β-mercaptoethanol (βME) lowered but did not eliminate the effect of DTT on activity: comparing 0 and 10 mM DTT assay conditions, NATL2A TalB prepared with βME increased in activity 40%, whereas NATL2A TalB prepared without βME increased in activity 130%. Notably, the same specific activity (37 $\pm$ 2 U/mg) could be achieved in both preps, supporting the reversible nature of this effect.

Initially, the reversible effect of DTT on TalB activity appeared to indicate a possible role for redox-active thiols in TalB. There are three conserved cysteines in TalB that are not conserved in TalC (Figure 2-3). These are, however, conserved in *E. coli* and other transaldolase orthologs, although *E. coli* (not cloned in pET100) shows no DTT effect. Based on threading models with *E. coli* and human transaldolase, two of these cysteines (Cys-156 and Cys-170 in MED4 TalB) lie 8–10 Å apart on an adjacent α helix and β strand near the active site. It is possible that in the actual structure, in some conformations, they are close enough to form a disulfide bond. Site-directed mutagenesis of Cys-170 to Ser in MED4 TalB, however, failed to eliminate the DTT effect (Figure 2-10). Further, the DTT effect could not be reproduced in MIT9312 TalB, which was cloned in p15TvLic. The same specific activity was achieved for MIT9312 TalB with (40.2 U/mg) or without (40.1 U/mg) 10 mM DTT. Thus, for the same reasons cited above for the F6P effect, we believe the DTT effect is an artifact of the pET100 N-terminal His-tag.

Comparative kinetics of *Prochlorococcus* and phage transaldolases

The kinetic parameters of TalB from three *Prochlorococcus* strains (NATL2A, MED4, and MIT9312) and TalC from three cyanophages (P-SSM2, P-SSM4, and P-SSP7) were determined using the transaldolase assay. The results are summarized in Table 2.3. The average turnover number of cyanophage transaldolase is about one-third that of *Prochlorococcus* transaldolase (Figure 2.3). Michaelis constants for both F6P and E4P are similar in TalC and TalB. For both F6P and E4P, $k_{\mathrm{cat}}/K_{\mathrm{m}}$ of cyanophage TalC was about one-third that of *Prochlorococcus* TalB. Values of $k_{\mathrm{cat}}/K_{\mathrm{m}}$ for both substrates and both enzymes were low relative to the limit of diffusion (10^5–10^6 s^{-1} mM^{-1}) (Berg et al. 2007). Similarities in kinetic parameters seem to be at odds with a model in which phage transaldolase possesses kinetic advantages over host transaldolase.

Temperature- and pH-dependent activities of *Prochlorococcus* and phage transaldolases

Given that kinetic parameters were not dramatically different between phage and host transaldolase and did not seem to account for the use of TalC by phage, we were curious if there were differences between TalB and TalC under changing physiological conditions, such as temperature or pH. The effect of temperature on the two transaldolase types was therefore

Table 2.3: Specific activities, catalytic constants, Michaelis constants, and specificity constants of transaldolases from *Prochlorococcus* and cyanophages. All assays were done at 25°C in 50 mM Gly-Gly (pH 8.0), 15 mM $MgCl_2$, 200 μM NADH, 0.6 U triosephosphate isomerase, 0.06 U glycerol-3-phosphate dehydrogenase, 0.1–10.0 mM F6P, and 0.01–1.00 mM E4P. *Prochlorococcus* TalB assays also contained 10 mM DTT. For each enzyme and substrate, 2–6 separate experiments were carried out and the data fit to Equation 2.3 (methods), then the replicates averaged with propagated standard error.

	Specific activity	k_{cat}	K_m (mM)		k_{cat}/K_m (s^{-1} mM^{-1})	
	(μmol min^{-1} mg^{-1})	(s^{-1})	Fructose 6-P	Erythrose 4-P	Fructose 6-P	Erythrose 4-P
Prochlorococcus						
NATL2A TalB	22.5 ± 1.1	15.2 ± 0.7	1.1 ± 0.2	0.11 ± 0.02	13.6 ± 2.5	134 ± 29
MED4 TalB	21.6 ± 0.6	14.9 ± 0.4	1.5 ± 0.2	0.15 ± 0.02	9.8 ± 1.2	103 ± 14
MIT9312 TalB	31.5 ± 2.0	20.8 ± 1.3	1.0 ± 0.1	0.10 ± 0.04	20.8 ± 2.3	206 ± 75
Cyanophage						
P-SSM2 TalC	7.9 ± 0.7	3.6 ± 0.3	0.7 ± 0.1	0.08 ± 0.01	5.4 ± 0.7	48 ± 9
P-SSM4 TalC	12.4 ± 0.8	5.6 ± 0.4	1.3 ± 0.2	0.20 ± 0.05	4.5 ± 0.9	29 ± 7
P-SSP7 TalC	12.9 ± 0.4	5.9 ± 0.2	1.6 ± 0.2	0.07 ± 0.01	3.6 ± 0.5	86 ± 18

examined. The protocol was similar to other transaldolase assays except the reagents were preheated to the appropriate temperatures for 5 min before mixing to initiate the assay. As shown in Figure 2-11, there is an increase in transaldolase activity as temperature increases for TalB from *Prochlorococcus* MED4 and TalC from cyanophages P-SSP7 and P-SSM4. Interestingly, MED4 TalB increases more dramatically in activity from 25°C to 30°C than either of the two phage TalCs. However, the two phage TalCs continue to increase in activity all the way to 40°C, whereas the activity of MED4 TalB declines rapidly as the temperature is increased beyond 30°C. Thus, we did observe differences in the activities of TalB and TalC as a function of temperature in that TalC appears to be more thermally stable at high temperatures than TalC.

The effect of pH on the two transaldolases was also examined. The protocol was similar to other transaldolase assays except the buffer pH was varied from 5.5–10.0 in the assays, done using the Ultramark. The results are shown in Figure 2-12. MED4 TalB and P-SSP7 TalC have similar pH-rate profiles, with maximal activity around pH 7.0–7.5 and activity decreasing to approximately 50% of maximum at pH 10.0.

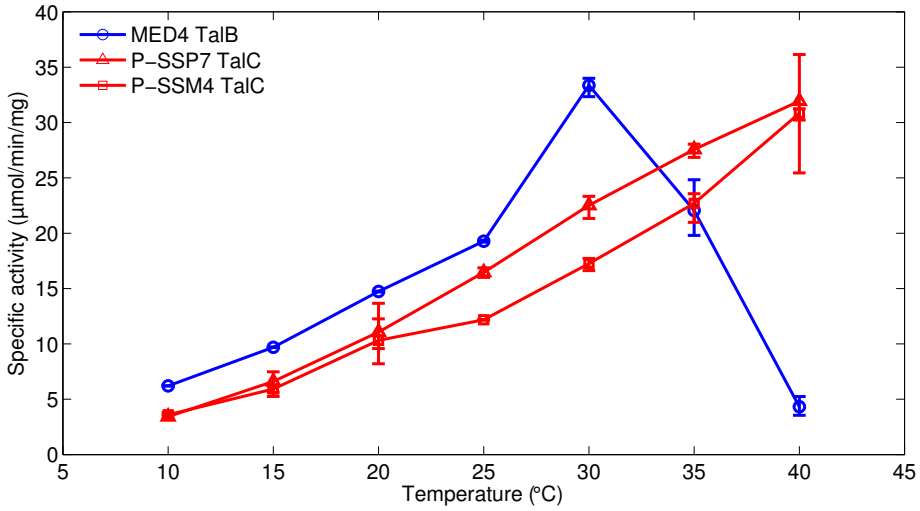


Figure 2-11: Temperature–rate profiles of *Prochlorococcus* TalB and cyanophage TalC. Error bars represent the maximum and minimum of three replicate assays.

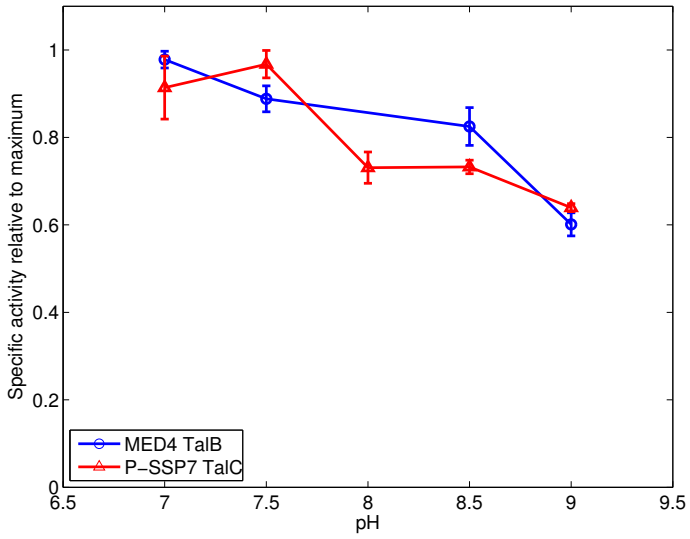


Figure 2-12: pH–rate profiles of *Prochlorococcus* TalB and cyanophage TalC. Error bars represent the standard deviation of three replicate assays.

Oligomerization state of *Prochlorococcus* and phage transaldolases

In other organisms, the quaternary structure of TalC is a decamer (dimer of pentamers) in solution, whereas the quaternary structure of TalB is a monomer or dimer (Samland and Sprenger 2009). The molecular weights in solution of *Prochlorococcus* TalB and cyanophage TalC were measured by comparing their retention times on a Superose 12 SEC column (optimal separation range 1–300 kDa) relative to molecular weight standards spanning 1.35–670 kDa (Figure 2-13). A standard curve resulting from a plot of log(molecular weight standard) versus retention time gave apparent molecular weights of *Prochlorococcus* TalB homo-oligomers and cyanophage TalC homo-oligomers.

Prochlorococcus MIT9312 TalB eluted as a single peak with mass of ∼35 kDa (Figure 2-13). This peak corresponds to the predicted molecular weight of a TalB monomer (40 kDa), including the His-tag (see Table 2.2). Cyanophage P-SSP7 TalC eluted as a single peak with mass of ∼133 kDa (Figure 2-13). This peak corresponds to the predicted molecular weight of a TalC pentamer (136 kDa), including the His-tag (see Table 2.2).

Crystal structure of *Prochlorococcus* MIT9312 TalB

The three-dimensional structure of full-length *Prochlorococcus* MIT9312 TalB was determined to 1.90-Å resolution (Figure 2-14). Molecular replacement found only one molecule in the asymmetric unit, and after refinement of the model, the quaternary structure server PISA (Krissinel and Henrick 2007) predicted that this protein is monomeric based on the absence of a large packing interface between any two molecules in the crystal. The resulting model contains a nearly complete chain for residues 0–332 (residue 0, a glycine, is a cloning artifact), with only the C-terminal residue (residue 333) missing. The model shows excellent geometry, with all residues in the favored and additional allowed regions of the Ramachandran plot. Data collection and refinement statistics are given in Table 2.4.

Prochlorococcus MIT9312 TalB consists of a single domain, an eight-stranded α/β barrel (Figure 2-14). Six parallel β strands (β_1–β_5 and β_8) form the core of the barrel. This core is surrounded by eight α helices (α_1–α_8) running approximately antiparallel to the β strands. There are six additional α helices (α_A–α_F), three of which are inserted in loop regions between β strands and α helices of the barrel: two are after β_2 (α_B and α_C) and one is after β_6 (α_D). The three remaining helices lie at the ends of the protein, with α_A at the N

Table 2.4: Data collection and refinement statistics for the *Prochlorococcus* MIT9312 TalB structure (PDB accession code 3HJZ).

Data collection	
Space group	$P2_12_12_1$
Cell dimensions	
a, b, c (Å)	42.8, 80.4, 97.9
Wavelength (Å)	1.54178
Resolution (Å)	40.19–1.9 (1.97–1.9)
R_{merge} (%)[a]	0.083 (0.372)
$I/\sigma I$	22.25 (5.7)
Completeness (%)	96.1 (68.3)
Redundancy	7.0 (5.5)
Refinement	
Resolution (Å)	40.19–1.95
No. reflections	26218
R_{work} (%)[b]	15.6
R_{free} (%)[c]	20.3
No. atoms	
Protein	2723
Water	341
Other	45
B-factors (Å^2)	
Overall	16.5
Protein	15.1
Water	25.9
Other	35.0
r.m.s. deviations	
Bond lengths (Å)	0.017
Bond angles (°)	1.45
Ramachandran plot	
% in most favored regions	93.1
% in additionally allowed regions	6.9
% in generously allowed or disallowed regions	0

[a] $R_{\text{merge}} = \Sigma \mid I - < I > \mid / \Sigma I$.
[b] $R_{\text{work}} = 100 \times \Sigma \mid F_{\text{obs}} - F_{\text{calc}} \mid / \Sigma F_{\text{obs}}$, where F_{obs} and F_{calc} are the observed and the calculated structure factors, respectively.
[c] R_{free} is calculated using 5% of total reflections randomly chosen and excluded from the refinement.

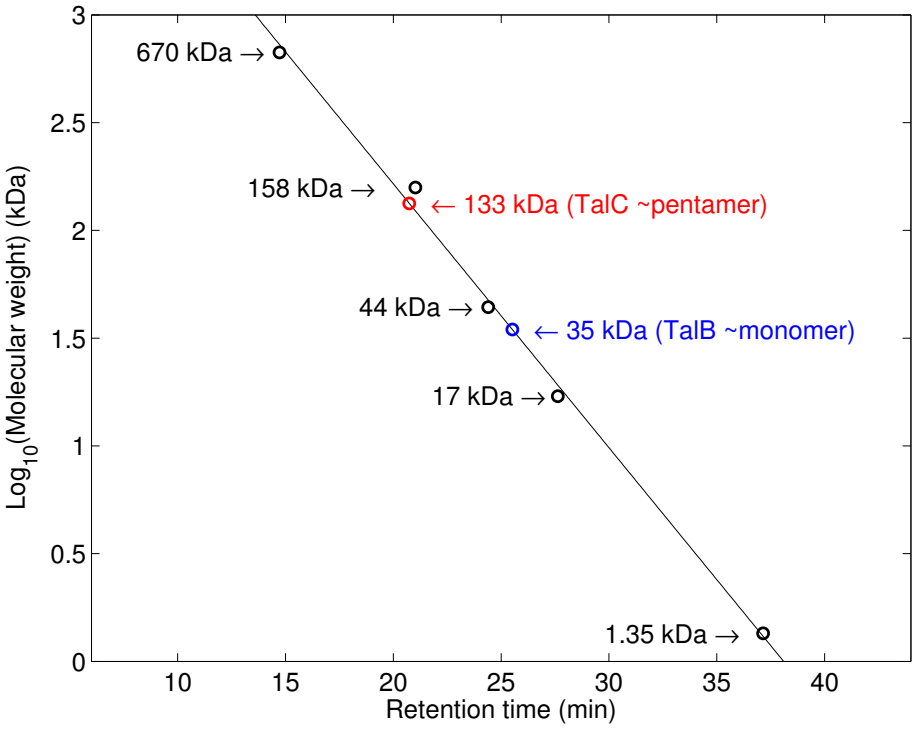


Figure 2-13: Molecular weight of *Prochlorococcus* MIT9312 TalB and cyanophage P-SSP7 TalC determined by SEC. Retention times for TalB are shown in blue, TalC in red, and molecular weight standards in black.

terminus and α_E and α_F at the C terminus. Helix α_F traverses and partly covers the active site at the C-terminal end of the barrel. A long loop wrapping around the barrel connects the last helix of the barrel (α_8) to the second to last C-terminal helix (α_E, which packs against α_4 and α_5 of the barrel).

The active site is located at the C-terminal ends of the β strands, and the walls of the active site space are formed by the loops connecting these β strands with the α helices. Although helix α_F runs across one half of the active site space, this space appears to be accessible from the bulk solution. The conserved Lys-135 (β_4), which likely forms a Schiff base with the substrate, is located at the bottom of the active site (Figure 2-16b). Near this residue are the conserved residues Asp-17 in β_1 and Glu-99 in β_3 (Figure 2-16b), which are likely involved in proton transfer during catalysis (Miosga et al. 1993). Also in this region is Phe-181 (Figure 2-16b), which has been implicated in substrate recognition, since mutation from Phe to Tyr changes substrate specificity in the human and *E. coli* enzymes from transaldolase (requiring E4P as acceptor substrate) to fructose-6-phosphate aldolase

(requiring no acceptor substrate) (Schneider et al. 2008) (Figure 2-8).

Homology model of cyanophage P-SSP7 TalC

A homology model of cyanophage P-SSP7 TalC (Figure 2-15) was built using structural information and the SWISS-MODEL package, with the structure of *T. maritima* TalC as the template structure (see methods). Homology models of the other two cyanophage TalC sequences and the other two *Prochlorococcus* TalB sequences were built in similar fashion with *T. maritima* TalC and *Prochlorococcus* MIT9312 TalB as template structures, respectively. The differences among cyanophage TalC models and among *Prochlorococcus* TalB structures/models were minimal, and therefore only the P-SSP7 and MIT9312 structures are discussed here.

The homology model of cyanophage P-SSP7 TalC reveals it also forms an eight-stranded α/β barrel (Figure 2-14). Six parallel β strands (β_1–β_5 and β_8) form the core of the barrel. This core is surrounded by eight α helices (α_1–α_8) that run approximately antiparallel to the β strands. There are two additional α helices (α_B and α_C): α_B is inserted between α_5 and α_6 of the barrel, whereas helix α_C lies outside and perpendicular to the core structure and is likely involved in subunit oligomerization.

The active site of cyanophage P-SSP7 TalC is similar to that of *Prochlorococcus* MIT9312 TalB. The positions of the Schiff-base-forming Lys-84, the proton-transferring Asp-6 and Glu-99, and the specificity-conferring Phe-130 are all in similar orientations (Figure 2-16b), as discussed below.

Structural comparison of *Prochlorococcus* and phage transaldolases

The structure of *Prochlorococcus* MIT9312 TalB (9312TalB) and homology model of cyanophage P-SSP7 TalC (PSSP7TalC) were aligned and superimposed using UCSF Chimera. The alignment of the two structures is shown in Figure 2-16. The subunits (Figure 2-16a) align well over the inner core of the α/β barrel. The six parallel β strands that constitute the β ladder of the core are positioned closely in the two structures. Surrounding the β ladder, 9 α-helices (α_1–α_8 and α_D/α_B) are also positioned closely, constituting the outer core of the α/β barrel. Strands β_6 and β_7 are not predicted in either cyanophage TalC or *Prochlorococcus* TalB, whereas they are observed in *E. coli* TalB and FsaA, human TALDO1, and *T. maritima* TalC.

The most striking difference between the two structures is the arrangement of exterior α helices. *Prochlorococcus* TalB (Figure 2-16b, blue) has helices α_B and α_C and a long extension of α_2. It also has the smaller exterior helix α_E followed by the very long helix α_F, which folds over the top of the subunit. None of these helices are present in cyanophage TalC. TalC is far more compact, with the significant exception of helix α_C, which protrudes far away from the core structure (Figure 2-16b, lower right). This helix has been implicated in subunit oligomerization by inter-subunit helix swapping in the decameric structure of *E. coli* Fsa (Thorell et al. 2002) and is also observed in the structure of *T. maritima* TalC. This is thought to be the major subunit interaction promoting oligomerization, in which two pentamers combine to form a decamer (Thorell et al. 2002).

A close-up of the aligned structures, showing the active sites, is shown in Figure 2-16b. In both structures, the positions of the catalytic lysine, glutamate, and aspartate residues are conserved. The role of lysine in formation of a Schiff base with the C2 keto group of the substrate is well established. Glutamate acts as a general acid-base in Schiff base formation, and aspartate acts as a general acid-base in carbon-carbon bond cleavage (Samland and Sprenger 2009). The position of phenylalanine is also conserved. This phenylalanine plays a role in maintaining transaldolase specificity; mutation of this phenylalanine to tyrosine in *E. coli* TalB or human TALDO1 changes specificity from transaldolase to F6P aldolase (Schneider et al. 2008). The observation of a conserved active-site geometry is consistent with the observation that both enzymes have transaldolase specificity.

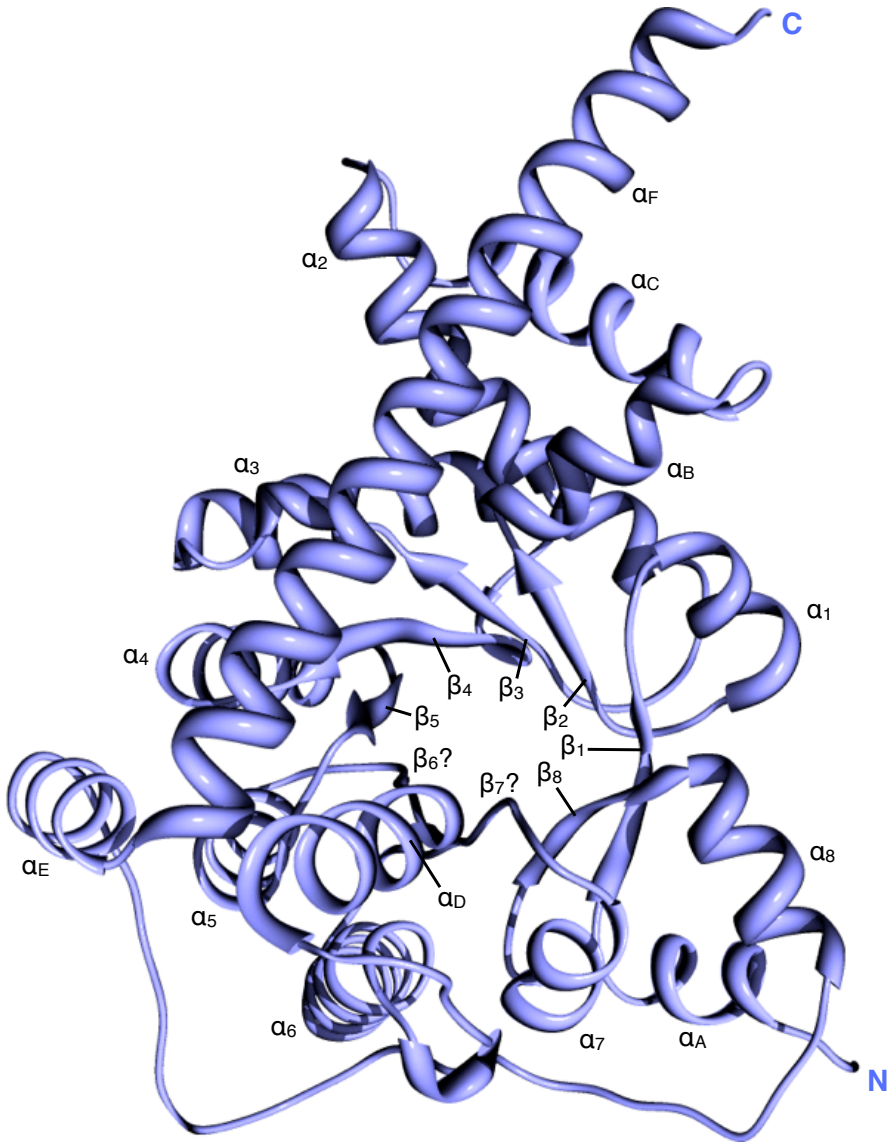


Figure 2-14: Structure of *Prochlorococcus* MIT9312 TalB subunit. Naming of α helices and β strands follows Figure 2-3a and the assignments made in Thorell et al. (2002).

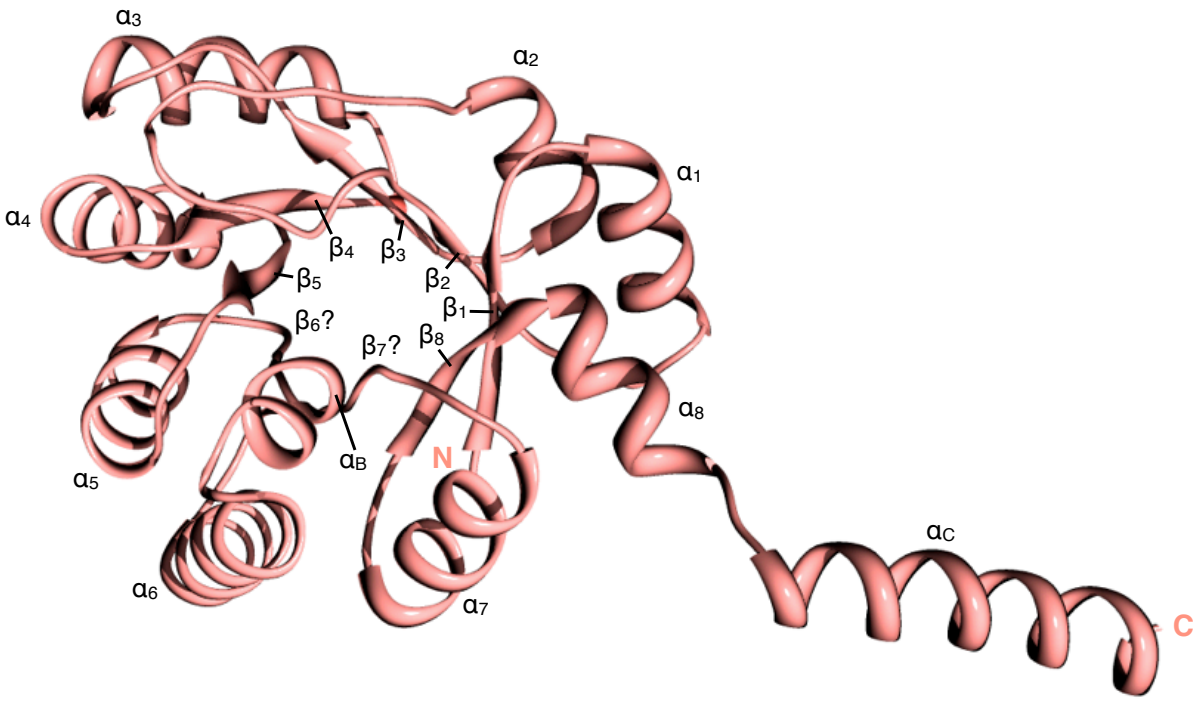


Figure 2-15: Homology model of cyanophage P-SSP7 TalC subunit. Naming of α helices and β strands follows Figure 2-3a and the assignments made in Thorell et al. (2002).

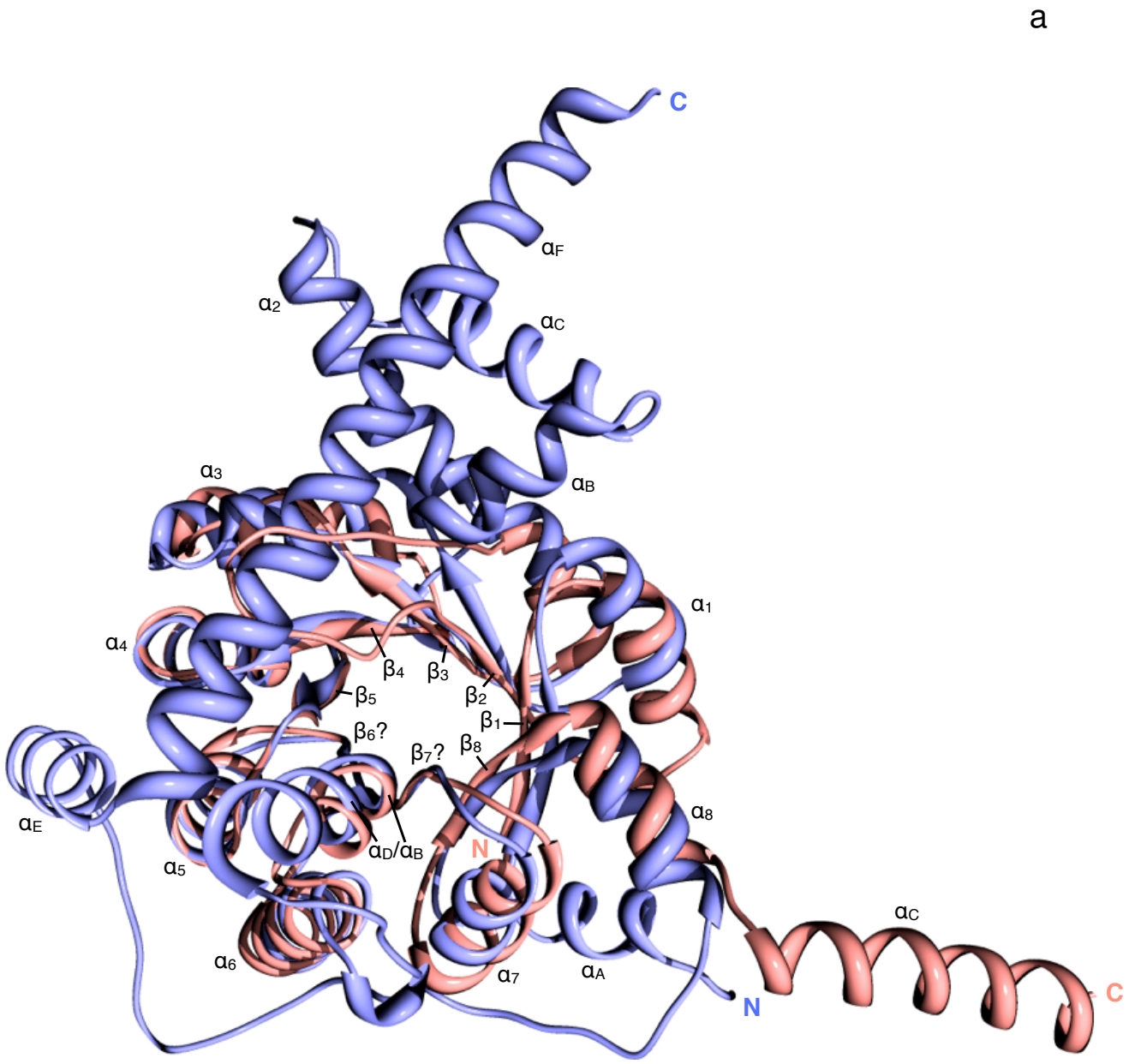


Figure 2-16: Superimposed structures of cyanophage TalC and *Prochlorococcus* TalB. Cyanophage P-SSP7 TalC was modeled on TalC from *T. maritima* (PDB accession code 1VPX); *Prochlorococcus* MIT9312 TalB was solved by x-ray crystallography (PDB accession code 3HJZ). (a) Subunits of cyanophage P-SSP7 TalC model (pale red) and *Prochlorococcus* MIT9312 TalB structure (pale blue) superimposed, showing conserved α/β-barrel core and variation in exterior arrangement of helices and loops. Naming of α helices and β strands follows Figure 2-3a and the assignments made in Thorell et al. (2002).

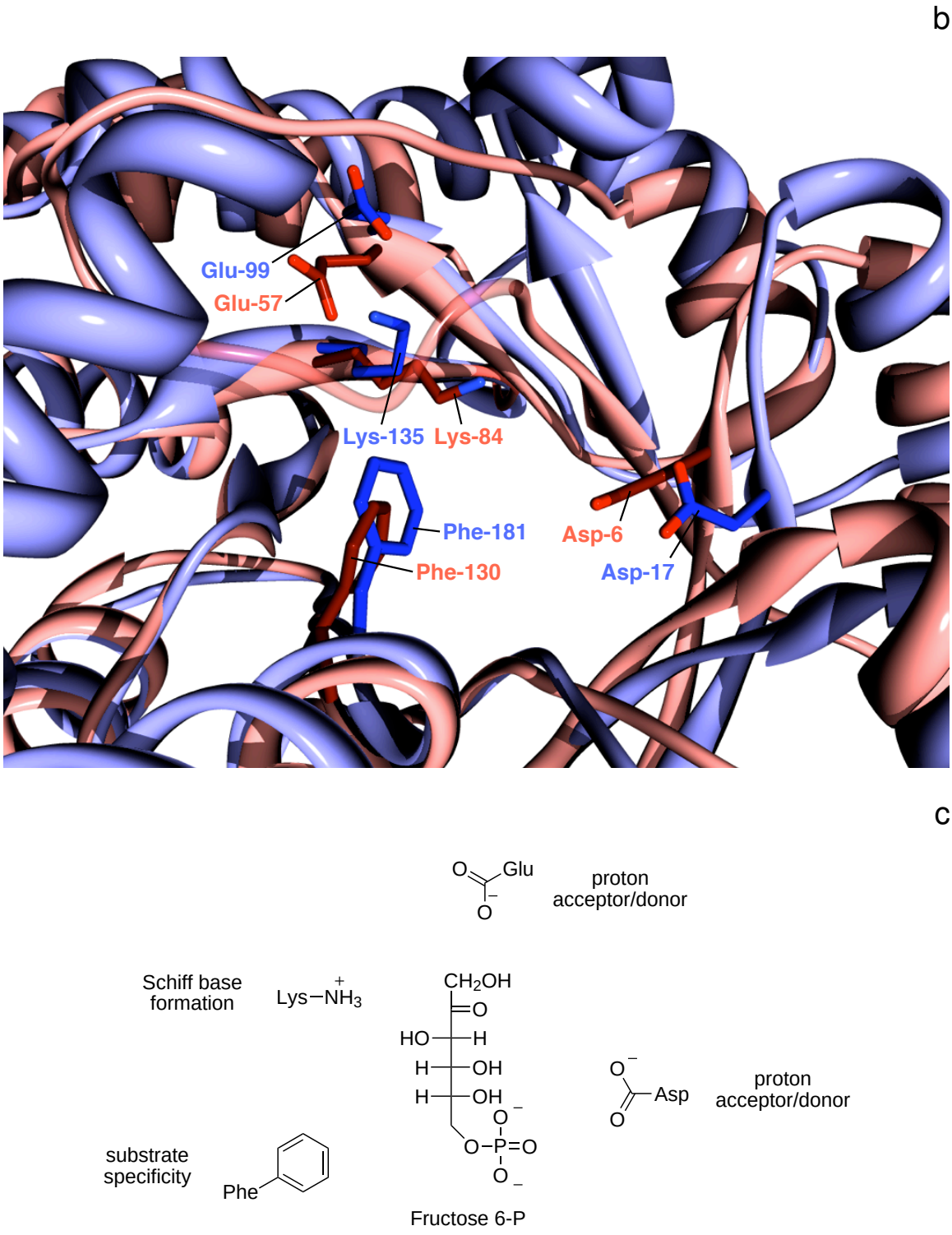


Figure 2-16: (continued) (b) Active sites of cyanophage P-SSP7 TalC model (pale red) and *Prochlorococcus* MIT9312 TalB structure (pale blue) superimposed, showing conserved arrangement of key active-site residues, with P-SSP7 TalC and MIT9312 TalB numbering colored dark red and dark blue, respectively. Oxygen atoms of glutamate and aspartate are colored bright red; nitrogen atoms of lysine are colored bright blue. (c) Position of these key residues relative to the substrate F6P and their roles in catalysis.

Discussion

In this project, we set out to address why cyanophage infecting *Prochlorococcus* encode a transaldolase (TalC) different from the host transaldolase (TalB). We asked, are there kinetic differences between TalC and TalB that can explain why cyanophage encode TalC? In the course of this study, we had to overcome obstacles with the transaldolase assay, protein expression, and purification tag effects. We characterized the kinetic parameters of three cyanophage TalC proteins and their corresponding *Prochlorococcus* TalB proteins, i.e., from strains known to infect in the laboratory. No obvious kinetic advantage of TalC over TalB was apparent, however, from the kinetic parameters. Without an obvious kinetic explanation for the use of TalC by cyanophage, we have considered alternate explanations involving various metabolic pressures on *Prochlorococcus* and cyanophage.

Methodological obstacles

The coupled assay for transaldolase activity (Figure 2-2) monitoring NADH consumption is a standard assay, but problems were introduced by substrate impurities. Certain pentose phosphate pathway intermediates, such as E4P and S7P, are present at very low concentrations in the cell and as such must be synthesized chemically. There is, however, very little commercial demand for these substrates, so they can be expensive and difficult to obtain in homogeneous form. S7P is not available from Sigma-Aldrich, for example. E4P is available commercially, but its preparation from Sigma-Aldrich is only 60% pure. Literature from Sigma-Aldrich and our own assays indicated that approximately 3% by mass of this E4P preparation is actually GAP, which introduced background NADH consumption. This problem was overcome by consumption of GAP before starting the reaction with transaldolase. We were also able to precisely determine the concentrations of E4P and F6P in stock solutions using endpoint assays, which permitted accurate determination of substrate concentrations required for K_{m} measurements. An additional problem with the transaldolase assay—a long lag phase in certain TalB enzymes—was subsequently determined to be a purification tag artifact, and these artifacts are discussed below.

Obtaining soluble overexpressed *Prochlorococcus* TalB also proved to be a problem. Under standard expression conditions with BL21 Star (DE3) cells grown at 25°C and induced with IPTG, *Prochlorococcus* TalB was only moderately overexpressed, and there was

no detectable soluble protein. Under these same conditions, cyanophage TalC was highly overexpressed, and a significant fraction of this protein was soluble. Low solubility of overexpressed proteins in *E. coli* is often the result of improperly folded proteins, which form insoluble inclusion bodies (Fahnert et al. 2004). In an attempt to get soluble TalB, we tried multiple expression strains, growth temperatures, and IPTG concentrations. Soluble TalB was ultimately obtained using a low growth temperature (13°C) and an expression strain (ArcticExpress) optimized for this temperature. The cold-adapted GroEL/GroES chaperonins made by this strain are designed to help proteins fold at cold temperature, which itself slows down metabolism and protein synthesis, presumably allowing time for the proteins to fold. This appears to have been the mechanism behind the synthesis of soluble TalB in this strain.

Two striking features were observed in *Prochlorococcus* TalB synthesized from the pET100 expression vector, but these features could not be reproduced in *Prochlorococcus* TalB synthesized using a different pET vector, p15TvLic, a derivative of pET-15b. The first feature, a lag in activity before time-dependent turnover was achieved, was found to be abolished by preincubation of the TalB with a small amount of F6P. The second feature was the dependence of maximal TalB activity on reductant (DTT or βME). The lag was initially frustrating but then quite exciting when the 'F6P effect' was discovered. The lag in *Prochlorococcus* TalB activity was intriguing, and we initially wondered if this could have some physiological relevance. The concentration of F6P required to diminish the lag to a normal level (1 mM) is the same as the K_{m} of *Prochlorococcus* TalB for F6P (Table 2.3). Thus, the range in which this F6P effect is observed is within the range where *Prochlorococcus* TalB is at an appreciable fraction (around half) of its maximal activity. The DTT effect was also exciting, because other enzymes in the pentose phosphate pathway are known to be regulated by the cellular redox environment. The conserved cysteines in *Prochlorococcus* TalB, however, were the same as those in *E. coli* TalB, which exhibits no DTT effect. Additionally, there was no pair of cysteines in the structure that appeared capable of forming a disulfide bond. Finally, there is no precedent for thiol/disulfide regulation of transaldolases. Neither the F6P effect nor the DTT effect could be reproduced in *Prochlorococcus* TalB with a different N-terminal His-tag, and we therefore attributed these effects to the His-tag of the pET100 vector. However, the kinetic parameters of *Prochlorococcus* TalBs synthesized using pET100, with F6P preincubation and 10 mM DTT, closely match the kinetics of

TalB synthesized using p15TvLic. Thus, it seems likely that these extra steps were able to overcome the tag-associated artifacts.

Kinetic and structural comparison of phage and host transaldolases

Cyanophage TalC and *Prochlorococcus* TalB proteins share only 24–29% sequence identity, and additionally there are several large gaps in the TalC sequence relative to TalB. The structures of TalB and TalC reflect these differences, but they also hint that these two enzymes are in many respects quite similar. The major distinguishing feature of the TalC structure, its protruding C-terminal helix, appears to account for the pentameric oligomerization state of the native protein observed by SEC. Similarly, analysis of the TalB structure indicates a weak dimer interface, which helps explain the monomeric quaternary structure of its native protein observed by SEC. Although there are significant differences in the number and arrangement of exterior helices in the two structures, the core α/β-barrel and active-site structure is remarkably well conserved. Both structures have the conserved orientation of active-site residues that is found in other transaldolase structures.

We chose kinetic studies as the major method to determine whether TalC and TalB might have important functional differences and whether these differences might provide some explanation for cyanophage use of a non-host-like enzyme. What we found in most cases was that there were not significant differences between the kinetic properties of phage TalC and host TalB. The most striking difference was in their turnover numbers, with TalB ∼three-fold higher than TalC. Thus, contrary to what might have been expected based on our hypothesis, the phage enzyme does not have a higher activity than the host enzyme, and in fact the opposite is true. The K_{m}s are also very similar between TalB and TalC. Thus, the $k_{\mathrm{cat}}/K_{\mathrm{m}}$ was also ∼three-fold higher for TalB than for TalC. From this kinetic analysis, then, *Prochlorococcus* transaldolase exceeds cyanophage transaldolase in both rate and efficiency.

Thus, based solely on kinetics, there does not seem to be any advantage conferred by TalC. If we consider the metabolic costs of enzyme production, the situation gets a little better for TalC. Given that TalC is about two-thirds the molecular weight of TalB, it takes about two-thirds the amino acids and ATP to synthesize TalC. In this sense, specific activity (units per mg) may be a more appropriate metric than turnover number (units per active site). The specific activity comparison gives TalB a two-fold advantage over TalC. This is

still significant but not as dramatic as the difference in turnover number, and it is perhaps more relevant to a cost–benefit analysis of phage infection.

In regions of the ocean where *Prochlorococcus* and cyanophage are found, the surface temperature ranges from 5–29°C (Johnson et al. 2006). As such, we were curious if there were temperature dependences of phage and host transaldolase that might illuminate key differences in the phage enzyme relative to the host enzyme. The temperature–rate data suggest differential thermal stability of phage and host transaldolases. Cyanophage TalC increases in activity up to 40°C, whereas *Prochlorococcus* TalB activity declines sharply above 30°C. While these results are intriguing, they are difficult to interpret. Changes in temperature compound the difficulties with in vitro assays. Because it is impossible to accurately mimic the cellular milieu, including various molecluar chaperones that modulate responses to heat, it is difficult to interpret the meaning of in vitro assays done over fluctuating temperatures. However, even if there are differential thermal stabilities of phage and host transaldolases, it is unclear whether they would be relevant in vivo, as *Prochlorococcus* TalB loses activity only above 30°C, which is near the upper temperature limit of *Prochlorococcus* (Johnson et al. 2006). An alternative explanation for the temperature data is that thermal stability may be merely an indicator of overall stability. It could be that TalC is stable under a wider range of conditions than TalB, for example particular intracellular conditions induced by phage infection.

Little is known about the intracellular conditions of *Prochlorococcus*, such as pH, redox state, salt concentration, and metabolite concentrations. Research on plant chloroplasts, however, indicates that photosynthetic electron transport, coupled to proton flow into the thylakoid lumen, results in alkalization of the stroma (equivalent to the cytoplasm of cyanobacteria), which activates several Calvin cycle enzymes by approaching their pH optima (Blankenship 2002). Thus, there appears to be a link between photosynthetic activity, pH, and carbon metabolism, and we wondered whether TalB and TalC might have different pH optima. In our experiments, however, the pH–rate profiles show similar trends for the two enzymes. Nevertheless, it seems promising that TalB and TalC could be differentially affected by changing intracellular conditions. We simply may not know enough about the intracellular environment of *Prochlorococcus* to know what to look for.

Non-kinetic explanations for cyanophage use of TalC

If kinetic differences cannot account for cyanophage using TalC, what are the most likely alternative explanations? In considering such alternatives, it is useful to discuss the current state of our thinking regarding the role of TalC in cellular metabolism during infection. As part of the pentose phosphate pathway (Figure 2-1), transaldolase helps oxidize F6P to Ru5P and NADPH, with some of the Ru5P recycled to keep the cycle going. R5P, which can be easily formed from from Ru5P, and NADPH are key precursors to DNA nucleotides. DNA nucleotides, in turn, are critical to DNA replication. We propose that TalC and the pentose phosphate pathway are critical during infection to produce NADPH and R5P for phage genome replication. The critical position of transaldolase in the PPP has been established by several studies showing it as the rate-limiting step in the non-oxidative portion of the PPP (Heinrich et al. 1976, Banki et al. 1996). The importance of the PPP to cyanophage, furthermore, is buttressed by the presence in several cyanophage genomes of genes for glucose-6-phosphate dehydrogenase and 6-phosphogluconate dehydrogenase (Weigele et al. 2007, Millard et al. 2009, Sullivan et al. ress), the two NADP-reducing enzymes of the PPP.

The question still remains though, if transaldolase is important, why don't cyanophage just encode TalB like the host? It is much more likely for a phage to acquire its host's gene than for it to acquire an exogenous one. Perhaps the simplest answer is that phage encode TalC by an accident of history, having acquired the *talC* gene by horizontal gene transfer from a non-cyanobacterial host or from some other horizontal gene transfer event. Whatever its origin, TalC provided the necessary flux through the transaldolase reaction, and even if the enzyme was not as efficient as the host enzyme, there was no need for the phage to improve on something that was good enough. We find this possibility unconvincing, however, since there are so many opportunities for phage–host gene transfer. It's likely that if the host gene were just as good, we would have found some phages with the host gene, but we haven't. Assuming there are differences between TalB and TalC that can explain the maintenance of TalC by cyanophage, we consider two possible explanations. The first concerns pressures at the gene and genome level of the phage. The second concerns pressures at the protein and proteome level of the host.

One possible explanation follows from the fact that because TalC is shorter than TalB, its gene is also shorter and therefore takes up less space in the host genome. The average size of

cyanophage *talC* is ∼650 bp, compared to the average size of *Prochlorococcus*/*Synechococcus talB* of ∼1,060 bp. Given that cyanophage genomes are as small as 45,000 bp, a 1,000-bp gene takes up a significant fraction of a phage genome. There are hard upper limits on genome size in phages because the phage capsid has a finite size. The compact nature of *talC* may provide an advantage to cyanophage by allowing them to carry other genes with the genome space saved. Any kinetic disadvantage of the TalC protein may be insignificant compared to the gained advantage in genome flexibility. This argument is supported by a similar trend in the size of other non-host-like metabolic cyanophage genes. Genes for the PPP enzymes mentioned above, glucose-6-phosphate dehydrogenase (*zwf*) and particularly 6-phosphogluconate dehydrogenase (*gnd*), are distinct from the host orthologs, and the phage genes in both cases are smaller. Phage *gnd* averages 1,030 bp whereas *Prochlorococcus*/*Synechococcus gnd* averages 1,420 bp, and phage *zwf* averages 1,440 bp whereas *Prochlorococcus*/*Synechococcus zwf* averages 1,520 bp. A more detailed analysis is necessary to determine whether this trend is statistically significant, but it seems the evolutionary advantage of getting by with a more compact gene could be significant.

A second possible explanation comes from recent work by Waldbauer (2009) looking at the proteome of *Prochlorococcus* MED4 over the diel cycle. He showed that while some 80% of genes in this strain oscillated with the day–night cycle at the mRNA level, many genes had little or no apparent periodicity at the protein level. One notable exception was transaldolase. Of all the proteins in the pentose phosphate pathway and the closely associated Calvin cycle, only TalB changed in abundance more than two-fold over the diel cycle (Waldbauer 2009). TalB abundance showed a two-fold drop in the morning hours of the diel cycle, which, given the unchanging abundances of other proteins in the PPP, may be a critical factor in controlling flux through this pathway. If so, cyanophage-encoded TalC may provide a way around this regulation. If a cyanophage infects its host in the morning hours and requires flux through the PPP, TalB abundance may be insufficient for this purpose, allowing a key role for TalC. Additionally, if TalB abundance in the host is controlled by protein degradation as well as synthesis, this could explain why the phage uses TalC. Because it differs significantly in sequence from TalB, TalC may lack a degradation sequence found in TalB, such as an N-terminal protein degradation signal. Further work investigating the regulation of protein synthesis and degradation in *Prochlorococcus* is needed to better inform this hypothesis. For this and many other studies, the dynamics of mRNA and protein

levels over the course of infection will be a critical direction of future research.

Acknowledgments

We acknowledge the Midwest Center for Structural Genomics for structural work. We thank Allison Ortigosa, Daniela Hristova, Mohammad Seyedsayamdost, Jun Wang, Mimi Cho, Ellen Minnihan, Joey Cotruvo, Kenichi Yokoyama, and Yimon Aye for assistance with protein purification and analysis, Rachel Buckley and Karen Allen for discussions on protein structure and crystallization, and Georg Sprenger and Tim Soderberg for discussions on the transaldolase assay. We thank Jeffrey Palm for assistance in manuscript preparation. This work was supported by the Gordon and Betty Moore Foundation, the Department of Energy (GTL), the National Science Foundation (C-MORE), a NIH grant to J.S., and a NIH Training Grant to L.R.T.

CHAPTER 3

Viruses infecting marine cyanobacteria express a Calvin cycle inhibitor alongside light reaction and pentose phosphate pathway genes

Luke R. Thompson, Qinglu Zeng, Libusha Kelly, Katherine H. Huang,
Maureen L. Coleman, and Sallie W. Chisholm
(manuscript to be submitted)

Abstract

Marine cyanophage are known to carry and express photosynthesis genes, whose products are thought to boost host photosynthesis and help repair damaged photosystems during infection. Although photosynthetic electron transport is typically coupled to carbon dioxide reduction to glucose by the Calvin cycle, no sequenced cyanophages have Calvin cycle genes. Cyanophages do, however, encode enzymes in the pentose phosphate pathway (PPP), which is used by cells to oxidize glucose to generate NADPH and ribose. We hypothesize that the light reactions of photosynthesis and the PPP operate concurrently in infected cells, with net consumption of glucose, and the NADPH and ribose produced used to power phage replication. To address this hypothesis, we screened 3 new and 21 published *Prochlorococcus* and *Synechococcus* cyanophage genomes for carbon and energy metabolism genes. We measured transcription of these genes during infection of *Synechococcus* WH8109 by cyanophage Syn9. Although no Calvin cycle genes were detected, we found widespread incidence of a gene for the Calvin cycle inhibitor CP12. Three PPP genes were also widely distributed in these cyanophages: *zwf* (glucose-6-phosphate dehydrogenase), *gnd* (6-phosphogluconate dehydrogenase), and *talC* (transaldolase). *talC* and *cp12* were the most prevalent of these four genes in the 24 cyanophage genomes; this trend was mirrored in the Global Ocean Sam-

pling metagenome. PPP, photosynthesis, and DNA biosynthesis genes were co-expressed with T4-like early genes during phage infection of *Synechococcus*. Thus, phage-encoded proteins for all three pathways appear to play a role early in infection, working in concert.

Introduction

Cyanophage are viruses that infect marine or freshwater cyanobacteria (Padan and Shilo 1973). The numerically dominant genera of marine cyanobacteria, *Prochlorococcus* and *Synechococcus* (Partensky et al. 1999, Scanlan and West 2002), contribute a significant proportion of primary productivity across large regions of the world's oceans (Li et al. 1983, Vaulot et al. 1995), and phages that infect them can be readily isolated on cultured host strains (Waterbury and Valois 1993, Sullivan et al. 2003). These cyanophage are potentially important agents of host mortality (Suttle and Chan 1994, Sandaa et al. 2009) and host evolution through phage-mediated horizontal gene transfer (Coleman et al. 2006, Sullivan et al. 2006).

Some form of horizontal gene transfer is implicated by the presence in most cyanophage genomes of 'host genes', i.e., genes with greatest similarity to cyanobacterial or bacterial genes rather than to genes from other phage types. Most of these genes have proposed functions in host metabolism, but because they are in fact encoded in phage genomes, we have proposed the term 'auxiliary metabolic genes' (AMGs) as a more descriptive term for this gene set (Breitbart et al. 2007, Appendix E). They are thought to provide supplemental support to key steps in host metabolism of significance to phage, thereby fostering a more successful infection. The encoded functions of AMGs found in a particular phage type appear linked to the metabolism of its host. For example, several genes for photosynthesis and carbon catabolism are shared between T4-like and T7-like cyanophages (Sullivan et al. 2005), which otherwise have completely different genome structures. Conversely, T4-like phages or T7-like phages from hosts with different metabolisms and found in non-marine environments, such as the enteric bacterium *E. coli*, while sharing many structural and other genes with their counterparts from cyanobacteria, have none of the photosynthesis or carbon metabolism genes found in cyanophages (Sullivan et al. in press, Appendix G).

Among the AMGs found in cyanophage genomes, photosynthesis genes are remarkably widespread. Genes encoding proteins involved in the light reactions are frequently observed,

encoding many functions involved in the light-driven production of ATP and NADPH. Evidence suggests that cyanophage-encoded photosynthesis genes are functional and perform an important auxiliary role in host metabolism during infection. The best studied example is *psbA*, which encodes the photosystem II core protein D1. Phage *psbA* is expressed and yields protein during infection (Lindell et al. 2005, Clokie et al. 2006). It has been proposed that phage-encoded D1 and other light reaction proteins help maintain photosynthetic electron flow during infection, ameliorating photosystem damage and boosting host photosynthetic activity (Lindell et al. 2004, 2005, Clokie et al. 2006). The precise role of photosynthesis genes during infection, however, is still an open question. Interestingly, the photosynthesis genes common in cyanophage are exclusively light reaction genes; no Calvin cycle genes have been reported for cyanophage genomes.

When grown under a natural light–dark cycle and not infected by phage, *Prochlorococcus* expresses all photosynthesis genes (light reactions and Calvin cycle) together such that maximal mRNA is present at sunrise (Zinser et al. 2009). This allows much of the ATP and NADPH from photosystems II and I to feed directly into the Calvin cycle, which uses their energy and reducing power to convert carbon dioxide into sugar. Genes for glycogen synthesis are maximally expressed at the same time (Zinser et al. 2009), suggesting that much of the sugar produced by the Calvin cycle is stored as glycogen. At sunset, genes for glycogen degradation and the pentose phosphate pathway (PPP) are maximally expressed (Zinser et al. 2009). The PPP oxidizes glucose to produce NADPH and ribose (Wood 1986a), providing reducing equivalents and carbon skeletons for nucleotide biosynthesis, genes for which are expressed at the same time (Zinser et al. 2009). Thus, in uninfected cyanobacteria, these metabolic activities are out of phase: light reaction and Calvin cycle genes are co-expressed in the morning, and PPP and DNA biosynthesis genes are co-expressed in the evening. An important additional form of regulation between the Calvin cycle and PPP is the scaffolding protein CP12, which has been shown to inhibit Calvin cycle enzymes in cyanobacteria (Tamoi et al. 2005). CP12 is expressed at night in *Prochlorococcus* (Zinser et al. 2009), consistent with its functioning to direct carbon flux away from the Calvin cycle and toward the PPP at night.

Interestingly, cyanophage genomes sequenced to date carry as many as three PPP genes: *zwf* for glucose-6-phosphate dehydrogenase, *gnd* for 6-phosphogluconate dehydrogenase, and *talC* for transaldolase (Sullivan et al. 2005, Weigele et al. 2007, Millard et al. 2009). Among

these genes, only *talC* has been studied at the transcriptional level, in cyanophage P-SSP7 (Lindell et al. 2007). Like other T7-like phages, timing of gene expression in P-SSP7 largely follows gene order. Remarkably, *talC* does not follow this trend: despite being the last gene in the P-SSP7 genome, *talC* is expressed much earlier, with photosynthesis genes *psbA* and *hli* and DNA biosynthesis gene *nrdJ* (ribonucleotide reductase) (Lindell et al. 2007). It appears that this phage regulates its gene expression such that AMGs for photosynthesis, the PPP, and DNA biosynthesis are co-expressed, possibly enabling the pathways represented by these genes to work together in the infection process.

Two pieces of evidence, then, are striking. First, although light reaction genes are co-expressed with Calvin cycle genes in *Prochlorococcus*, cyanophages have no Calvin cycle genes. Second, in at least one cyanophage, light reaction genes are co-expressed with PPP and nucleotide biosynthesis genes, contrary to what is seen in the transcription of host genes of uninfected cells. This suggests that phage may be redirecting the flow of carbon and energy in host metabolism to their own specific ends, namely, to fuel biosynthesis for phage replication. More formally, we hypothesize that during infection, cyanophage-encoded light reaction and PPP proteins work in concert, generating reducing equivalents and energy, which are readily consumed by cyanophage-encoded nucleotide biosynthesis proteins to fuel phage replication. To begin to address this hypothesis, we first examined 3 new and 21 published cyanophage genomes to assess their cache of genes for the light reactions, the PPP, and DNA biosynthesis, and to confirm that they all lack Calvin cycle genes. We also searched the phage genomes for the Calvin cycle inhibitor CP12, which, given the hypothesis above, could be an important regulator of host carbon metabolism for phage. For one T4-like phage, we determined whether its AMGs are co-expressed during infection, consistent with coordinated functions. Finally, we examined the prevalence of these genes in marine metagenomic databases to determine if their distribution among cultured isolates reflected relative frequencies in wild cyanophage populations.

Materials & Methods

Sequences and gene annotation

Twenty-four genomes from cyanophages infecting *Prochlorococcus* and marine *Synechococcus* were included in the analyses (Table 3.3), three of which are being introduced for the first

time in this paper (Table 3.2). Complete genome sequences and annotations of published genomes were downloaded from Integrated Microbial Genomes (Markowitz et al. 2006) and GenBank (Benson et al. 2008). New cyanophage genomes reported in this study were sequenced using the method of Henn et al. (2010) and annotated using the annotation pipeline described by Sullivan et al. (in press, Appendix G).

We searched the 24 cyanophage genomes for each photosynthetic electron transport, Calvin cycle, PPP, and DNA biosynthesis gene using the genome annotations. To check for the possibility of uncalled or miscalled genes in the annotations, we also searched the 24 genomes using TBLASTN with default parameters, an E-value cutoff of 1e–5, and using all *Prochlorococcus* or *Synechococcus* Calvin cycle, PPP, photosynthetic electron transport, and DNA biosynthesis genes as queries. In cases of positive hits for genes not reported in the annotations, multiple sequence alignments were used to confirm the presence of key conserved residues.

Metagenomic analyses

The complete Global Ocean Sampling (GOS) database (Rusch et al. 2007), current as of August 2009, was downloaded from CAMERA (Seshadri et al. 2007). This database contains 9,893,120 sequences and 8,047,788,530 bp, for an average read length of 813 bp.

Each GOS read was blasted against a database containing 12,683 sequences representing marine bacteria and phage genomes to recruit each read to its closest identifiably homologous marine genome. This database contains genomes of sequenced marine isolates, including Gordon and Betty Moore Foundation Marine Microbial Initiative genomes, NCBI marine isolates, and cyanophages available from the Genbank and CAMERA databases as of November 2008. BLASTN parameters were selected to allow for as low as 65% identity between a read and a hit sequence and to permit gaps in the alignment: blastall -p blastn -r 5 -q -4 -e 1e-4 -z 3000000000 -F "m L" -X 150 -U T. Best hits for each read were extracted from the BLAST output. Paired ends for each read were collected and compared to ensure that the GOS paired end for each read was recruited to either the same genome or to any of the 24 cyanophage genomes.

GOS reads recruited to cyanophage genomes using the above method and having both paired ends with greatest similarity to a T4-like cyanophage were assigned to T4-like cyanophage gene clusters, defined by Sullivan et al. (in press, Appendix G). Gene clusters are

referred to as 'core' or 'non-core' depending on whether or not they were found in each of 16 available T4-like cyanophage genomes. BLASTX with default parameters was used to blast each T4-like cyanophage GOS read against the set of all T4-like cyanophage genes. Hits were filtered with an E-value cutoff of 1e–4 and bit score cutoff of 40; additionally, the top five hits were required to map to the same gene cluster, or if there were fewer than five members of a cluster then all hits were required to map to the same cluster. Hit counts for each gene cluster were plotted against average gene length of the cluster.

Infection of *Synechococcus* WH8109 by cyanophage Syn9

Synechococcus WH8109 was maintained in SN medium (Waterbury and Willey 1988) made with 75% filtered seawater from the Environmental Systems Laboratory, Woods Hole, MA, USA. Salts and metals for SN medium were from Sigma-Aldrich (St. Louis, MO, USA). Cultures were grown in a 'sunbox', a modified Percival Scientific (Boone, IA, USA) I-35LL plant growth chamber with a 24-h light–dark cycle consisting of 5 h of increasing light from 0–320 μE m^{-2} s^{-1}, 5 h of 320 μE m^{-2} s^{-1}, 4 h of decreasing light from 320–0 μE m^{-2} s^{-1}, and 10 h of dark (Zinser et al. 2009). Temperature was maintained at 24 $\pm$ 0.2°C. On the day of infection, 2 h before dark, log-phase *Synechococcus* WH8109 (1×10^8 cells mL^{-1} by flow cytometry) was infected with cyanophage Syn9 (3×10^8 infective phage mL^{-1} by most probable number (MPN) assay (Tillett 1987)), resulting in a multiplicity of infection (MOI) of 3. Three replicate infected cultures and three replicate uninfected control cultures of 1 L each were maintained. Uninfected controls were given spent medium instead of phage lysate. Both spent medium and phage lysate were filtered through 0.2-μm polycarbonate filters (Millipore, Billerica, MA, USA) prior to addition. Following addition of phage lysate or spent medium, bottles were transferred to constant light of 50 μE m^{-2} s^{-1}.

Samples were taken at regular intervals for RNA and genomic DNA (gDNA) quantification. For RNA, 1-mL samples were centrifuged at 15,000$\times g$ for 2 min at 4°C, the supernatant aspirated, and the cell pellet flash frozen in liquid nitrogen and stored at −80°C. For phage and host gDNA quantification, 100-μL samples were filtered with 0.2-μm polycarbonate filters. The filtrate was diluted 1:1000 for extracellular phage gDNA quantification. For intracellular phage and host gDNA quantification, the filter was washed with three 1-mL volumes of preservation solution (10 mM Tris-HCl, 100 mM EDTA, 500 mM NaCl, pH 8.0) and flash frozen; the cells were subsequently resuspended in 650 μL 10-mM Tris-HCl (pH

Table 3.1: qPCR primers used in this study.

Gene	Forward primer	Reverse primer
Cyanophage Syn9		
g61	5'-GGTTTGGGTATCAGGGAAGG-3'	5'-AACATCAGCACCACACATCG-3'
g43	5'-GAAGTTGGAGCCTTTCATCG-3'	5'-ACCTCACACCCTCACTGTCC-3'
g20	5'-AATTGAAATCCGCAATGAGC-3'	5'-CATAGCGGGATCCATTTCC-3'
g23	5'-AACCTACGAGCAAGCAGACG-3'	5'-ATTGCCTTCAGGTCTTGTGC-3'
psbA	5'-CGGTGGGTCACTTTTCTCG-3'	5'-CGACCGAAGTAACCATGAGC-3'
nrdA	5'-CTGGGCATTGGTTTTATTGG-3'	5'-CCTTTTTCCATTGCCATACG-3'
zwf	5'-TTCTCCATCGTCTGGATTGG-3'	5'-GCAATCCTGCTTCTTTGAGG-3'
gnd	5'-CTAAGGTGGCTGAGCTTTGG-3'	5'-ACAGCAGCGTGAACAGTCC-3'
talC	5'-CCCGAGCTTATTGCTACTGC-3'	5'-AATCTGCTGCCATACCAAGC-3'
cp12	5'-CATCGAAAAGCACATTCAGG-3'	5'-CCTCGCAGTAGAGCTCAAGG-3'
Synechococcus WH8109		
rnpB	5'-GCCGATCTCTTTGAGTGTCG-3'	5'-GCTCTTACCGCACCTTTGC-3'

8.0) by aggitation in a Mini-Beadbeater (BioSpec, Bartlesville, OK, USA), the supernatant heated to 95°C for 15 min, then diluted 1:100.

Quantitative PCR and RT-PCR

Primer design

qPCR primers were designed from the genomes of cyanophage Syn9 (Weigele et al. 2007) and *Synechococcus* WH8109 (GenBank) using Primer3 (Rozen and Skaletsky 2000) with a GC clamp of at least 2 bp, yielding products of 150–200 bp. Primers were designed such that homologs among the Syn9 and WH8109 genomes could be distinguished. Sequences are given in Table 3.1. Primers were tested using Syn9 and WH8109 gDNA and were shown to have specific and concentration-dependent amplification of target DNA.

RNA extraction, DNase treatment, and cDNA synthesis

Synechococcus cell pellets were resuspend in 100 μL 10-mM Tris-HCl (pH 8.0), 100 units RNase inhibitor (SUPERASE-In, Ambion, Austin, TX, USA), and 15,000 units lysozyme (Ready-Lyse, Epicentre, Madison, WI, USA), and incubated at 37°C for 30 min, after which 15,000 units additional lysozyme was added, followed by 30 min at 37°C. RNA was extracted from this lysate using the Mini RNA Isolation II Kit (Zymo Research Corp., Orange, CA, USA), and RNA was eluted with nuclease-free water. This RNA was treated with 6 units

Turbo DNase I (Ambion). cDNA was made from this DNase-treated RNA using the iScript cDNA Synthesis Kit (Bio-Rad, Hercules, CA, USA); prior to cDNA synthesis, the reaction mixture lacking reverse transcriptase was heated to 65°C for 5 min and then cooled on ice.

Quantitative PCR

gDNA or cDNA copies were quantified using the QuantiTect SYBR Green PCR Kit (QIAGEN, Valencia, CA, USA) with a LightCycler 480 Real-Time PCR System (Roche Diagnostics, Indianapolis, IN, USA). qPCR reactions contained 0.5 μM forward and reverse primers and approximately 0.5 ng μL^{-1} cDNA. The amplification reaction consisted of an initial activation step of 15 min at 95°C, followed by 50 cycles of 15 s at 95°C (denaturation), 30 s at 56°C (annealing), and 30 s at 72°C (extension), followed by extension for 5 min at 72°C, followed by a melting curve from 50–90°C. Threshold cycle (C_T) of amplification was determined by the second derivative maximum method. Concentrations of phage and host gDNA over the time course were determined with standard curves of log(concentration of standard) versus C_T. Relative copy numbers of each cDNA over the time course were determined by the $\Delta\Delta C_T$ method (Pfaffl 2001), with *rnpB* (RNA component of ribonuclease P) as the internal calibrator gene (Fernández-González et al. 1998).

Results & Discussion

AMG content of all sequenced marine cyanophages

To assess the potential of cyanophage to intervene in particular host metabolic processes, we conducted a targeted search of all available cyanophage genomes for the presence of AMGs involved in photosynthetic electron transport, the Calvin cycle, the PPP, and nucleotide biosynthesis. Twenty-four sequenced genomes from cyanophages infecting *Prochlorococcus* and *Synechococcus* are currently available, including three new genomes of T7-like podoviruses introduced here (Table 3.2). This analysis allowed the identification of possibly overlooked AMGs and provided a concise overview of the genetic potential of cyanophage with respect to these metabolic pathways (Table 3.3).

Consistent with previous investigations of cyanophage genomes (Sullivan et al. 2005, Mann et al. 2005, Weigele et al. 2007, Millard et al. 2009), we observed photosynthetic electron transport genes *psbA* and *psbD* (photosystem II D1 and D2 proteins), *petE* (plas-

Table 3.2: Properties of three T7-like podovirus genomes from this study.

		Source water properties			Genome properties		
Strain	Original host	Location	Depth	Date	Size (kbp)	ORFs	%GC
P-RSP5	*Pro.* NATL1A	Red Sea, 29°28'N 34°55'E	130 m	13 Sep 2000	47.7	66	38.7
P-HP1	*Pro.* NATL2A	Hawai'i, 22°45'N 158°00'W	25 m	8 Mar 2006	47.5	64	39.9
P-SSP2	*Pro.* MIT9312	Sargasso Sea, 31°48'N 64°16'W	120 m	28 Sep 1995	45.9	56	37.9

tocyanin), *petF* (ferredoxin), PTOX (plastoquinol terminal oxidase), and *hli* (high-light inducible protein) (Table 3.3).

Although many cyanophage genomes contain photosynthesis genes, it is not clear whether they have been explicitly examined for the presence of Calvin cycle genes. This is an important consideration, as it has been postulated that the activity of phage-encoded photosystem II genes during infection leads to significant carbon fixation in the oceans (Sharon et al. 2007). We therefore examined these genomes, looking for Calvin cycle genes, using BLAST (see methods). We found no evidence of any Calvin cycle genes in these genomes, including those genes that are shared between the Calvin cycle and the PPP. However, the gene for an inhibitor of the Calvin cycle, CP12, was found, which has been a key link for understanding the role of cyanophage AMGs in manipulating host metabolism.

PPP genes were prevalent in these genomes, as has been documented previously (Sullivan et al. 2005, Mann et al. 2005, Weigele et al. 2007, Millard et al. 2009). *zwf*, *gnd*, and *talC* were found in 6, 8, and 20 of the 24 genomes, respectively (Table 3.3 and Figure 3-1). *zwf* and *gnd* encode glucose-6-phosphate dehydrogenase and 6-phosphogluconate dehydrogenase, respectively, which generate NADPH in the oxidative portion of the PPP. *talC* encodes transaldolase in the non-oxidative portion of the PPP. *zwf* and *gnd* were first identified in cyanophage Syn9 (Weigele et al. 2007) and are shown here to be found only in *Synechococcus* T4-like cyanophages (Table 3.3). *talC* was the first PPP gene to be identified in cyanophage (Millard et al. 2004, Sullivan et al. 2005) and is more widespread than *zwf* or *gnd*, as it is found in both T4-like and T7-like cyanophages (Table 3.3). The distinct presence/absence patterns across phage types (Table 3.3) are discussed below.

DNA biosynthesis genes were also observed, again consistent with previous studies (Sullivan et al. 2005, Mann et al. 2005, Weigele et al. 2007, Millard et al. 2009). Many of these cyanophage genomes carry *nrdAB* or *nrdJ* (class Ia or II ribonucleotide reductase),

Strain	Type	Original host	Location isolated	Size (kbp)	%GC	Reference	Pentose phosphate pathway				Photosynthetic electron transport						Nucleotide biosynthesis		
							talC	*cp12*	*gnd*	*zwf*	*psbA*	*psbD*	*petE*	*petF*	PTOX	*hli*	*nrd*	*cobS*	*thyX*
P-SS2	Sipho	*Pro.* MIT9313	Atlantic slope waters	107.5	52.3	c		×									*J*		×
P-SSP7	T7	*Pro.* MED4	Sargasso Sea	45.0	38.8	f	×				×					×	*J*		
P-SSP2	T7	*Pro.* MIT9312	Sargasso Sea	45.9	37.9	This study	×				×					×	*J*		
P-RSP5	T7	*Pro.* NATL1A	Red Sea	47.7	38.7	This study	×				×					×	*J*		
P-HP1	T7	*Pro.* NATL2A	Hawai'i	47.5	39.9	This study	×	×			×					×	*J*		
P60	T7	*Syn.* WH7803	Georgia coastal river	47.9	53.2	h											*J*		×
Syn5	T7	*Syn.* WH8109	Sargasso Sea	46.2	55.0	e											*J*		×
P-HM2	T4	*Pro.* MED4	Hawai'i	183.8	38.0	a	×	×			×	×				×	*AB*	×	×
P-HM1	T4	*Pro.* MED4	Hawai'i	181.0	38.0	a	×	×			×	×			×	×	*AB*	×	×
P-SSM4	T4	*Pro.* NATL2A	Sargasso Sea	178.2	36.7	f	×	×			×	×			×	×	*AB*	×	×
P-RSM4	T4	*Pro.* MIT9303	Red Sea	176.4	38.0	a	×	×			×				×	×	*AB*	×	×
P-SSM7	T4	*Pro.* NATL1A	Sargasso Sea	182.2	37.0	a	×	×			×			×		×	*AB*	×	×
P-SSM2	T4	*Pro.* NATL1A	Sargasso Sea	252.4	35.5	f	×	×			×		×	×		×	*AB*	×	×
S-PM2	T4	*Syn.* WH7803	English Channel	196.3	37.8	g					×	×				×	*AB*	×	×
S-SSM7	T4	*Syn.* WH8109	Sargasso Sea	232.9	39.0	a	×	×			×		×		×	×	*AB*	×	×
Syn33	T4	*Syn.* WH7803	Gulf Stream	174.4	40.0	a	×	×			×	×	×	×		×	*AB*	×	×
S-ShM2	T4	*Syn.* WH8102	Atlantic shelf waters	179.6	41.0	a	×	×	×		×	×	×			×	*AB*	×	×
Syn1	T4	*Syn.* WH8101	Woods Hole	191.2	41.0	a	×	×	×		×	×	×			×	*AB*	×	×
S-SM1	T4	*Syn.* WH6501	Atlantic slope waters	178.5	41.0	a	×	×	×	×	×	×	×		×	×	*AB*	×	×
S-SSM5	T4	*Syn.* WH8102	Sargasso Sea	176.2	40.0	a	×	×	×	×	×	×	×		×	×	*AB*	×	×
Syn9	T4	*Syn.* WH8012	Woods Hole	177.3	40.5	d	×	×	×	×	×	×	×		×	×	*AB*	×	×
Syn19	T4	*Syn.* WH8109	Sargasso Sea	175.2	41.0	a	×	×	×	×	×	×	×		×	×	*AB*	×	×
S-RSM4	T4	*Syn.* WH8103	Red Sea	194.5	41.1	b	×	×	×	×	×	×	×	×	×	×	*AB*	×	×
S-SM2	T4	*Syn.* WH8017	Atlantic slope waters	190.8	40.0	a	×	×	×	×	×	×	×	×	×	×	*AB*	×	×

Table 3.3: Distribution of pentose phosphate pathway, photosynthetic electron transport, and select nucleotide biosynthesis genes in 24 cyanophage genomes. Phage strains are classified by phage type (Sipho, siphovirus; T7, T7-like podovirus; T4, T4-like myovirus) and host of isolation (*Pro.*, *Prochlorococcus*; *Syn.*, *Synechococcus*). Gene name abbreviations: *J*, *nrdJ*; *AB*, *nrdAB*. References: [a]Sullivan et al., 2010; [b]Millard et al., 2009; [c]Sullivan et al., 2009; [d]Weigele et al., 2007; [e]Pope et al., 2007; [f]Sullivan et al., 2005; [g]Mann et al., 2005; [h]Chen and Lu, 2002.

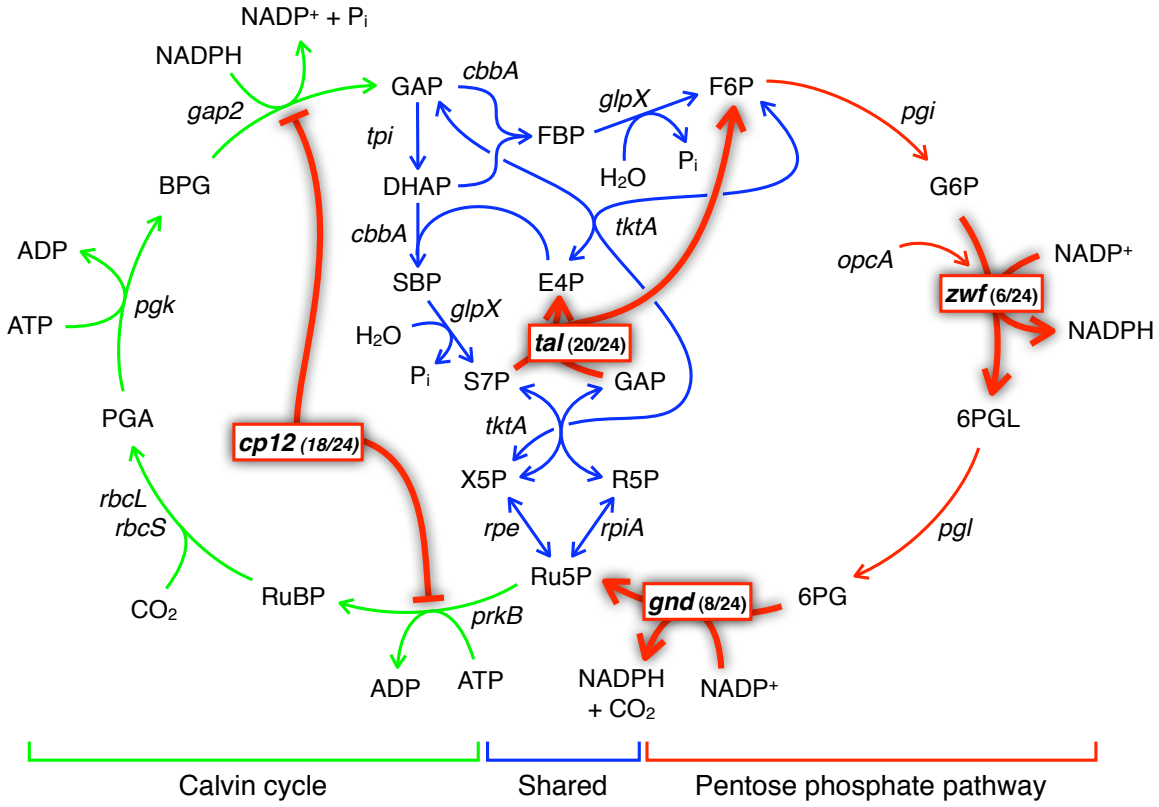


Figure 3-1: Diagram of the pentose phosphate pathway (PPP) and the Calvin cycle in cyanobacteria, showing which genes/enzymes are specific for the PPP (red), specific for the Calvin cycle (green), or shared between the two pathways (blue). Genes found in cyanophage are in bold with thick lines and denoted with the number of genomes out of 24 in which they are found. CP12 (*cp12*) is grouped with the PPP since it shuts off the competing Calvin cycle, inhibiting phosphoribulokinase (*prkB*) and glyceraldehyde-3-phosphate dehydrogenase (*gap2*). Glucose-6-phosphate dehydrogenase (*zwf*, EC 1.1.1.49) oxidizes glucose 6-phosphate to 6-phosphoglucono-lactone, generating NADPH. 6-phosphogluconate dehydrogenase (*gnd*, EC 1.1.1.44) oxidizes 6-phosphogluconate to ribulose 5-phosphate, generating NADPH and carbon dioxide. Transaldolase (*tal*, EC 2.2.1.2) reversibly transfers a three-carbon dihydroxyacetone moiety from sedoheptulose 7-phosphate to glyceraldehyde 3-phosphate, generating erythrose 4-phosphate and fructose 6-phosphate. Metabolite and gene/protein abbreviations are defined on page 23, and metabolite structures are given on page 146.

cobS (cobalt chelatase for adenosylcobalamin biosynthesis, an essential cofactor for the host ribonucleotide reductase, NrdJ (Stubbe et al. 2001)), *thyX* (thymidylate synthase, which converts dUMP to dTMP in pyridine biosynthesis), and several other pyrimidine and purine biosynthesis genes. Other than *thyX*, these pyrimidine and purine biosynthesis genes were sporadically distributed and much less common and therefore are not included in Table 3.3.

A gene encoding the Calvin cycle inhibitor CP12 was found in 18 of the 24 cyanophage genomes examined (Table 3.3 and Figure 3-1). We recently reported the presence of *cp12* in T4-like cyanophages (Sullivan et al. in press, Appendix G). Here we report that one T7-like cyanophage and one siphovirus also carry *cp12*. These reports represent the first identification of *cp12* in phage genomes, and this gene may prove to be a key link for understanding the role of host genes in cyanophage infection. CP12 is an intrinsically unstructured protein widespread in photosynthetic organisms. In plants and cyanobacteria, CP12 inhibits two enzymes in the Calvin cycle (phosphoribulokinase and glyceraldehyde-3-phosphate dehydrogenase), promoting flux through the PPP (Tamoi et al. 2005). It is particularly notable that *cp12* was found in all three types of cyanophages (T4-like myoviruses, T7-like podoviruses, and a siphovirus, as shown in Table 3.3). Besides *cp12*, only genes for ribonucleotide reductase, thymidylate synthase, and terminase have been documented in all three cyanophage types, but unlike *cp12*, each of those genes is also found in at least one non-cyanophage. Therefore, *cp12* appears to be not only a uniquely cyanophage adaptation but one that has been acquired independently by all three major lineages of cyanophages, underscoring a potentially central role in the cyanophage infection process.

The presence of genes for CP12 in all three major lineages of cyanophages, along with the absence of Calvin cycle genes and presence of light reaction, PPP, and DNA biosynthesis genes, suggests that cyanophage are exerting influence on host metabolism as follows: First, we propose that the AMGs are co-expressed during infection such that their products can act in concert, unlike the light-dependent expression patterns of host metabolic genes in uninfected cells. Second, we propose that the particular AMGs carried by cyanophage have been selected for to fill key metabolic bottlenecks that arise during infection. We present a model that incorporates our experimental data and the particular gene complement of cyanophage: the light reactions and PPP are activated, while the Calvin cycle is deactivated, producing reducing equivalents, energy, and carbon skeletons for use in phage DNA biosynthesis and replication.

Cyanophage CP12 and PPP genes are expressed with photosynthesis and DNA biosynthesis genes

Previous studies of gene expression during cyanophage infection have focused mainly on T7-like cyanophage P-SSP7 (Lindell et al. 2005, 2007), which lacks *zwf*, *gnd*, and *cp12*. We were curious about the regulation of gene expression in a cyanophage with a richer cache of PPP genes (i.e., *zwf*, *gnd*, and *talC* as well as *cp12*). We chose host–phage system *Synechococcus* WH8109 and T4-like cyanophage Syn9 for these studies because this phage carries all four of the AMGs of interest (three enzyme genes plus *cp12*) and its genome is published and highly curated (Weigele et al. 2007). Besides the four PPP genes (*zwf*, *gnd*, *talC*, and *cp12*), we measured expression of photosystem II gene *psbA* (D1 protein) and ribonucleotide reductase gene *nrdA* (alpha subunit) to see if they were co-expressed with the PPP genes. To reference the timing of expression of the AMGs to core phage genes central to phage replication, we also measured expression of T4-like phage early genes *g61* (DNA primase) and *g43* (DNA polymerase) and late genes *g20* (portal protein) and *g23* (major coat protein). Timing of gene expression in T4-like phages is controlled by promoters (early, late, and sometimes middle) and is not dependent on gene order across the genome (Miller et al. 2003b).

Phage adsorption at time zero was followed rapidly by host gDNA degradation (Figure 3-2a). After ∼2.5 h, phage gDNA increased inside host cells, and cells began to lyse ∼6.5 h after the infection began (Figure 3-2a). All cyanophage Syn9 genes being studied here were expressed during infection, with the T4-like early genes (Figure 3-2b–d). The transcripts of *psbA*, *nrdA*, and PPP genes, including *cp12* (Figure 3-2c–d), were detectable just 30 min after infection, and increased until ∼4 h after infection, at which point they leveled off and remained constant throughout the remainder of the latent period. Only the structural genes *g20* and *g23* (Figure 3-2b) were expressed later, first detectable at 90 min, increasing until 4 h, when expression also leveled off.

Using early genes *g61* and *g43* and late genes *g20* and *g23* as guides (Figure 3-2b), along with the infection time course (Figure 3-2a), we can see clearly that *zwf*, *gnd*, *talC*, *cp12*, *psbA*, and *nrdA* are all expressed early in the infective process (Figure 3-2c–d), consistent with their proposed functions. Early genes in T4-like phages are known to be involved in establishing infection and tend to encode enzymes or regulatory proteins, whereas late genes

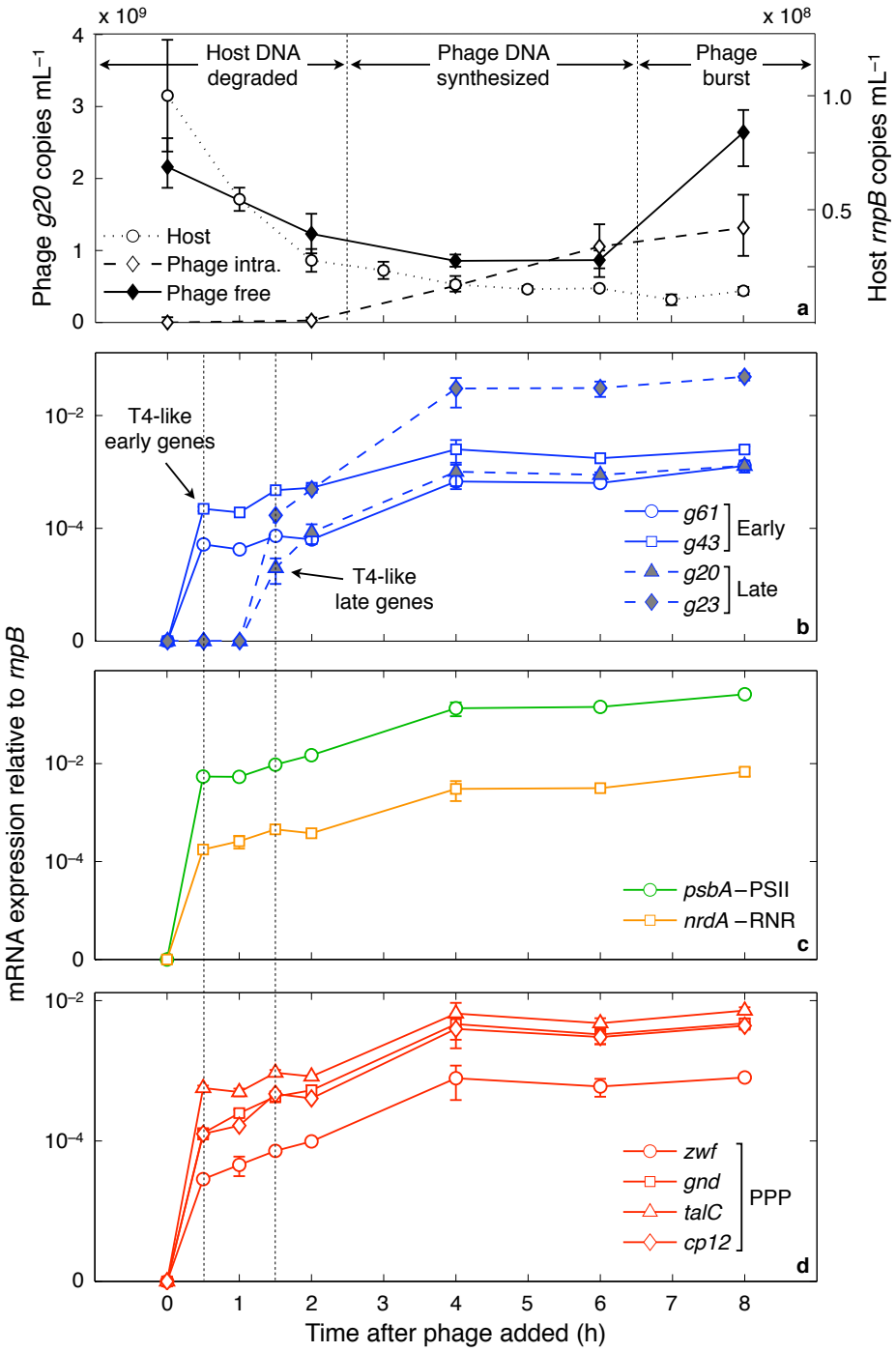


Figure 3-2: Infection dynamics and gene expression of cyanophage Syn9 infection of *Synechococcus* WH8109. (a) Free and intracellular phage *g20* copies mL^{-1} and host *rnpB* copies mL^{-1}, as proxies for genome copies mL^{-1}. (b) mRNA of T4-like early genes *g61* (DNA primase) and *g43* (DNA polymerase) and late genes *g20* (portal protein) and *g23* (major coat protein). (c) mRNA of photosystem II (PSII) D1 gene *psbA* and ribonucleotide reductase (RNR) gene *nrdA*. (d) mRNA of PPP genes *zwf* (glucose-6-phosphate dehydrogenase), *gnd* (6-phosphogluconate dehydrogenase), and *talC* (transaldolase) and *cp12* (Calvin cycle inhibitor CP12).

are involved in assembling progeny virions and tend to encode structural proteins (Birge 2006, Miller et al. 2003b, Roucourt and Lavigne 2009). Enzymes and regulatory proteins are required early in the infection cycle to take over host metabolism and produce DNA and protein, which form new phage virons later in the infection cycle. In cyanophage Syn9, PPP enzymes and CP12 produced early in infection would be positioned to direct glucose toward the production of NADPH and ribose for use in phage replication. Indeed, the AMGs we measured for photosynthesis, the PPP, and DNA biosynthesis were all expressed after just 30 min, several hours before phage gDNA was detected inside cells (Figure 3-2).

Further, like Lindell et al. (2007), we observed that transaldolase (*talC*) was co-expressed with photosynthesis (*psbA*) and ribonucleotide reductase (*nrdA*) genes (Figure 3-2c–d). The same was true for the other three PPP genes (*zwf*, *gnd*, and *cp12*). If the offset between transcription and translation of these genes is similar, we would expect the protein products of these AMGs to be present simultaneously in infected cells. If these enzymes and regulatory proteins are functional, we could imagine the following scenario: photosynthetic electron transport, aided by phage-encoded photosynthesis proteins, produces ATP and NADPH; the PPP, aided by phage-encoded enzymes, produces NADPH and ribose; the Calvin cycle, inhibited by phage-encoded CP12, does not consume ATP and NADPH; and DNA biosynthesis, aided by phage-encoded ribonucleotide reductase, consumes ATP, NADPH, and ribose, producing DNA nucleotides. The DNA nucleotides produced by this joint phage–host metabolism could be used for phage genome replication.

Comparative genomics and metagenomics suggest that cyanophage genes fill key metabolic bottlenecks

All four of the PPP genes found in marine cyanophage, including the Calvin cycle inhibitor gene *cp12*, were found in multiple phage genomes (Figure 3-1 and Table 3.3), with even the least frequent gene found in 6 of 24 genomes. Two of the genes, *talC* and *cp12*, were also found in multiple phage types, with *talC* in both T4-like and T7-like cyanophages and *cp12* in all three major cyanophage types (Table 3.3). Thus, rather than a sparse, sporadic distribution, we find these genes in multiple phage genomes and in some cases multiple phage types (e.g., T4-like myoviruses, T7-like podoviruses, siphoviruses), underscoring selective pressures for their maintenance in phage genomes. Although there are eleven genes involved in the PPP, only these four are carried by cyanophages examined to date (Figure 3-1). They

likely represent important steps in the PPP, i.e., metabolic bottlenecks during infection.

From the perspective of phage evolution, which AMGs are carried by cyanophages is the product of selective pressures: the cost of maintaining a gene in a size-limited genome versus the benefit of encoding a novel metabolic function. We might postulate that since phage genome size is dictated by capsid size and smaller genomes cannot carry as many genes, the genes carried by smaller genomes should represent the most strongly selected metabolic functions; similarly, across phages of the same type and similar size, those genes found more frequently should be more strongly selected for. T7-like genomes are the smallest of the three types (45.0–47.7 kbp, versus 107.5 kbp for siphovirus P-SS2 and 174.4–252.4 kbp for T4-like cyanophages; see Table 3.3), and they contain *talC* and *cp12*. Across T4-like cyanophages, by far the best sampled group of marine cyanophages, *talC* and *cp12* are also the most prevalent, found in 16 of 17 genomes, with *gnd* and *zwf* found in 8 and 6 genomes, respectively (Table 3.3). Based on the frequency and diversity of genomes in which they are found, *talC* and *cp12* appear to be under stronger selection than *gnd* and *zwf* and therefore may be more critical to maintaining PPP flux under infection.

To expand our understanding of the selective pressures on cyanophage PPP genes from culture to wild populations, we analyzed data from the Global Ocean Sampling (GOS) expedition, which has so far yielded over 8 gigabases of marine metegenomic sequence (Rusch et al. 2007). A significant portion of this sequence data is not cellular but viral in origin, a sort of 'metagenomic bycatch' that may result from intracellular viruses or viruses stuck to cells or collection filters. Much of this viral sequence is from T4-like cyanophages (Williamson et al. 2008). We searched for all T4-like cyanophage genes, differentiating them based on whether they are found in all T4-like cyanophages ('core') or a subset of T4-like cyanophages ('non-core'), as defined by Sullivan et al. (in press, Appendix G). In particular, we were interested whether we could find significant incidence of cyanophage PPP genes in the environment and whether their relative proportions could be correlated with presence/absence trends seen in sequenced genomes.

There was a direct proportionality between the size of T4-like cyanophage core genes and the number of putative T4-like cyanophage sequence reads observed in the GOS database (red circles in Figure 3-3). This proportionality—i.e., the number of reads counted increasing as a function of gene size—is what one expects to see for genes that are found in all T4-like cyanophage genomes (Sullivan et al. in press, Appendix G), confirming their core

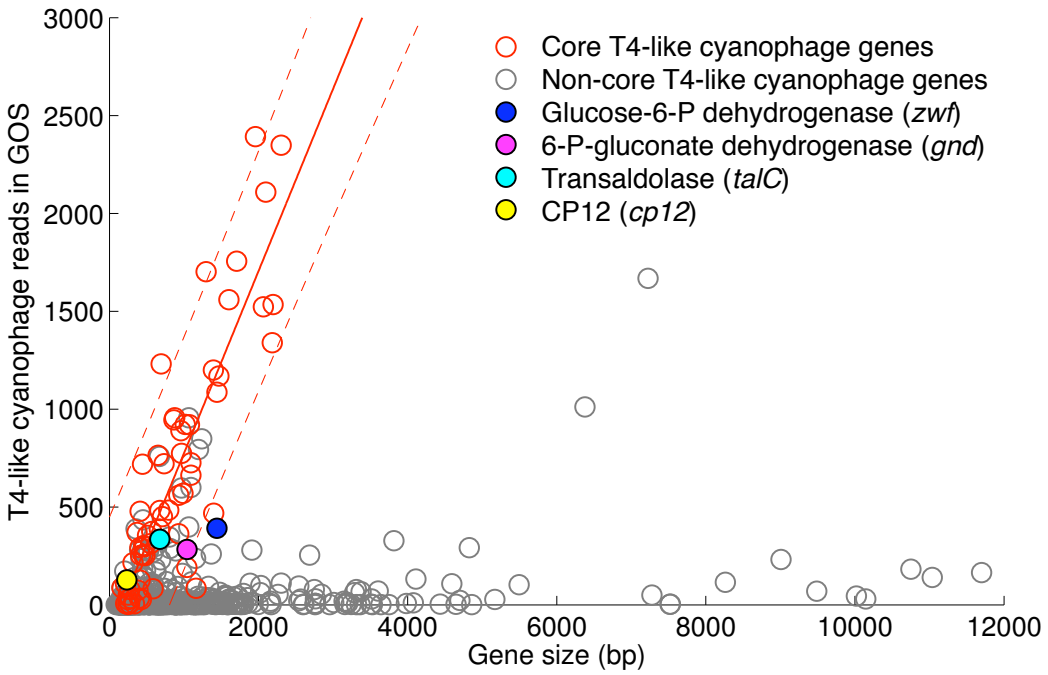


Figure 3-3: Abundance of cyanophage *cp12* and pentose phosphate pathway genes relative to core and other non-core T4-like cyanophage genes in the surface ocean. Genes are colored based on whether they are found in all T4-like cyanophage genomes (core, red) or found in less than all T4-like cyanophage genomes (non-core, gray); *zwf*, *gnd*, *talC*, and *cp12* are non-core and are filled in with bright colors. For each gene found in a T4-like cyanophage, the number of times that gene was observed as a sequence read (i.e., the 'read count') in the Global Ocean Sampling (GOS) database is plotted versus gene size. This type of plot reveals that, because the DNA fragments cloned and sequenced are constant, larger genes tend to be cloned and sequenced more frequently than smaller genes; genes having the same copy number in a sample should produce a linear plot of read count versus gene size. The linear regression (with 95% confidence interval) of this pattern is shown for core gene clusters. GOS reads were counted only if they had greatest similarity to a gene in a cyanophage genome (see methods). For clarity, eight non-core genes between 12000–24000 bp are not shown. Definitions of core and non-core are defined in Sullivan et al. (in press, Appendix G).

status. Genes found only sporadically in T4-like cyanophage genomes should lie well below the regression line, as we see for some non-core genes (gray circles in Figure 3-3), whereas genes found in most T4-like cyanophage genomes should lie near the regression line, as we see for other non-core genes (gray circles in Figure 3-3). All four of the PPP-associated cyanophage genes were observed in GOS (filled-in circles in Figure 3-3). *talC* and *cp12* fell with the core genes (within a 95% confidence interval), consistent with their presence in all but one T4-like cyanophage genome. *gnd* and *zwf*, found in less than half of T4-like genomes, were less abundant: *gnd* fell right on the 95% confidence interval for core genes, and *zwf* fell outside this boundary; these lower abundances in GOS are consistent with their less frequent occurrence among the sequenced genomes. Thus, remarkably, for these four genes, the prevalence trend observed in the sequenced genomes is borne out in GOS.

It is striking that with the limited number of phage genomes available, the relative frequncies of these genes in the cultured database and in the wild are in such agreement. It is possible that the concordance between prevalence patterns is influenced by the relative abundance of sequences used to recruit reads from the GOS database. This possibility is under investigation at the time of this writing.

Increased selection for *talC* and *cp12* relative to other PPP genes, as inferred from relative frequency of occurrence in genomes, suggests that these proteins could be limiting during infection. Indeed, previous studies have shown transaldolase to be the rate-limiting step in the non-oxidative portion of the PPP (Heinrich et al. 1976, Banki et al. 1996). Further, recent work by Waldbauer (2009) provides evidence that transaldolase might be limiting under certain conditions in *Prochlorococcus*. He examined protein levels over the diel cycle of *Prochlorococcus* MED4. Many proteins did not show significant oscillations in protein abundance over the diel cycle, even though most of them had significant day–night oscillations at the mRNA level. Of all the proteins in PPP and Calvin cycle, only transaldolase (TalB) changed significantly in abundance over the diel cycle, dropping two-fold in the morning hours (Waldbauer 2009). Given the constant abundances of the other PPP and Calvin cycle proteins, this fluctuation in TalB abundance could be a critical factor in controlling flux through the PPP over the natural growth cycle of *Prochlorococcus*. In light of these data, let us assume that the transaldolase reaction limits the PPP during the day because of a low amount of host TalB protein. If infection were to occur in the morning, when TalB is limiting, TalC expression from cyanophage could circumvent this bottleneck

imposed by host protein levels. If transaldolase is the keystone step in the PPP, this helps explain why *talC* is found in more cyanophage genomes than any other PPP gene.

CP12: a hidden clue to metabolic fluxes during infection

The *cp12* gene is distinct from *talC*, *zwf*, and *gnd* in that it encodes a regulatory protein rather than an enzyme. Whereas phage transaldolase, glucose-6-phosphate dehydrogenase, and 6-phosphogluconate dehydrogenase are proposed to increase flux through the PPP directly, CP12 is proposed to increase flux through the PPP indirectly, via inhibition of the Calvin cycle. Here, the issue is not so much metabolic bottlenecks during infection as much as competition with another pathway, since the PPP and Calvin cycle are linked (Figure 3-1). Cyanophage *cp12* is a unique AMG with respect to both how its protein works and how its protein is regulated.

To our knowledge, CP12 is the first phage-encoded protein with the putative ability to interact directly with enzymes in central carbon metabolism, as most known phage regulatory proteins target transcription, translation, replication, or bacterial defense mechanisms (Roucourt and Lavigne 2009). Further, most of these known phage regulatory proteins are unique to phage (Roucourt and Lavigne 2009), whereas CP12 was likely acquired from a cyanobacterial host.

CP12 is also notable for how it is known to be regulated in host systems. Activity of CP12 is controlled by the redox state of the cell and relative levels of the nicotinamide adenine dinucleotide cofactors NAD, NADH, NADP, and NADPH (Wedel et al. 1997). Most other proteins encoded by phage AMGs are tacitly assumed to be constitutively active during infection, although there is evidence that cyanobacterial glucose-6-phosphate dehydrogenase is regulated by light and redox (Gleason 1996, Sundaram et al. 1998, Hagen and Meeks 2001). In *Arabidopsis thaliana*, CP12 is activated by oxidizing conditions and a low NADP(H)/NAD(H) ratio (i.e., decreased NADP and NADPH relative to NAD and NADH), which lead to formation of an intramolecular disulfide in CP12 and ternary complex formation of CP12 with PRK and GAPDH (Marri et al. 2008). The day–night cycle of cyanobacteria provides a helpful framework for thinking about CP12 regulation, since NAD(P)(H) levels fluctuate over the diel cycle (Tamoi et al. 2005) and the PPP is upregulated and the Calvin cycle downregulated at night (Stöckel et al. 2008, Zinser et al. 2009). Indeed, *Prochlorococcus cp12* is maximally transcribed at sunset, putatively leading

to maximal protein levels at night (Zinser et al. 2009). Onset of darkness in cyanobacteria leads to a more oxidizing environment but also to a decrease in the NADP(H)/NAD(H) ratio, activating CP12 (Tamoi et al. 2005). This has the effect of inhibiting the Calvin cycle at night, when the cell is harvesting energy from glucose via the PPP and cannot have the competing Calvin cycle running. If there is a similar decrease in the NADP(H)/NAD(H) ratio under infection, then phage-encoded CP12 could likewise be activated, stimulating carbon flux through the PPP.

The function of CP12 therefore presents a clue to metabolic fluxes during cyanophage infection (Figure 3-4). If the Calvin cycle is turned off by phage-encoded CP12, then the energy (ATP and NADPH) produced by the light reactions is not used to fix carbon dioxide: phage infection decouples the light reactions from the Calvin cycle. Instead, two pathways that under the light–dark cycle are offset by 12 h—the light reactions of photosynthesis and the PPP—are potentially occurring simultaneously in the host cell. If photosynthetic energy (ATP and NADPH) is not used by the Calvin cycle, and additional energy (NADPH) is produced by the PPP, then the most likely use for this ATP and NADPH is to fuel phage nucleotide biosynthesis, particularly via the phage-encoded ribonucleotide reductase, as summarized in Figure 3-4. The suggestion that cyanophage photosynthesis proteins lead to a net increase in primary production (Sharon et al. 2007) is an intriguing one. The data presented here, however, indicate that the Calvin cycle is likely inhibited by CP12 during infection and that glucose is likely oxidized to generate further reducing equivalents. According to our model (Figure 3-4), photosynthetic energy is useful to infecting phage but does not lead to new carbon fixation; rather, it is used to power phage replication from existing reduced carbon.

Stoichiometry of phage replication

We have argued here that the predominant metabolic push in the infected host, as driven by the expression of phage AMGs, is toward the synthesis of DNA for phage genome replication. This is not an obvious conclusion, however. Phage are composed of both DNA and protein, and both must be synthesized for phage replication. If cyanophage direct host metabolic flux toward increased synthesis of DNA relative to protein, this implies that the demand for DNA relative to protein is greater for phage than for *Prochlorococcus*. It also implies that the DNA in the host chromosome, even if it is completely digested to single nucleotides, is

insufficient to supply all of the DNA in progeny phage.

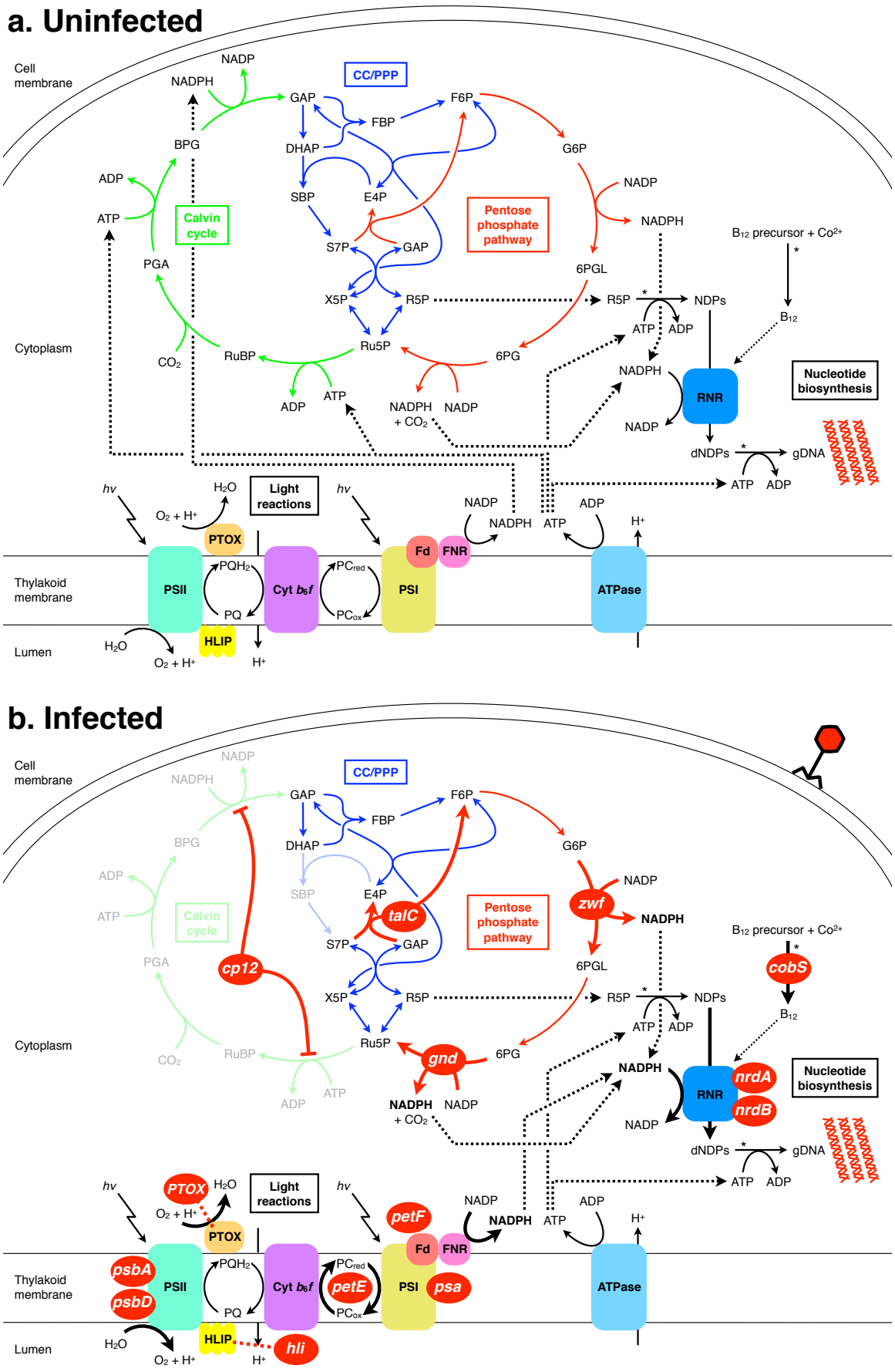
To gauge the relative fluxes of DNA and protein biosynthesis required for phage and host replication, we estimated the stoichiometric ratios of protein to DNA in a typical cyanophage and in a typical *Prochlorococcus* cell. For a typical *Prochlorococcus* MED4 cell, we assumed a total protein content of 1.85×10^8 amino acids and an average amino acid mass of 109 Da (both calculated using estimates from Waldbauer (2009)), for a total mass of 20 GDa protein. The MED4 genome is 1.66 Mbp (Rocap et al. 2003), for a total mass of 1 GDa DNA. The estimated protein/DNA ratio of a MED4 cell is thus ∼20. For a typical cyanophage, we used the coliphage T4, which is structurally similar to T4-like cyanophages; the T4 genome (169 kbp) is slightly smaller than the genomes of sequenced T4-like cyanophages (174–252 kbp), but the protein/DNA ratios should be similar. Using ultrastructural information about the copy number of each phage protein in a T4 virion (Leiman et al. 2003), we calculated a total mass of 110 MDa protein. Including terminal redundancy, the 169-kbp T4 genome forms a chromosome of 172 kbp (Leiman et al. 2003), which corresponds to 106 MDa. The estimated protein/DNA ratio of T4 and therefore a T4-like cyanophage is thus ∼1. As we can see, the protein/DNA ratio of the host (∼20) is twenty times the protein/DNA ratio of phage (∼1). The metabolic fluxes, therefore, required to replicate phage are biased toward synthesizing DNA over protein, relative to the normal metabolic fluxes in the host for its own replication.

Estimates of the number of phage produced from each *Prochlorococcus* cell (burst size) likewise support the hypothesis that de novo nucleotide production is critical to phage replication. T4-like cyanophage burst sizes range from 40 (Brown et al. 2006) to ∼150 (L. R. Thompson and Q. Zeng, unpublished results); for this discussion we will assume an average burst size of 100 phage/cell. To produce 100 phage with a chromosome size of 200 kbp requires 20 Mbp of DNA, but the host *Prochlorococcus* chromosome is only 1.66 Mbp. Even is the burst size were only 10 phage/cell, there would still not be enough DNA from the host chromosome to produce that many phage. Without other sources of nucleotides for scavenging, replicating phage require the biosynthesis of new nucleotides. From the analyses presented here, therefore, it seems logical that phage would direct host metabolism to preferentially synthesize DNA, consistent with the metabolic gene content of cyanophages and with our model.

Acknowledgments

We want to thank Jake Waldbauer and Erik Zinser for valuable discussions on carbon metabolism and its regulation. We are grateful to Matthew Sullivan for providing phages for genome sequencing and Debbie Lindell for providing *Synechococcus* WH8109 gene sequences. Brianne Holmbeck and Sara Roggensack provided assistance in maintaining cultures. We thank Alexa Price-Whelan, Arne Materna, Katya Frois-Moniz, Sébastien Rodrigue, and Kolea Zimmerman for advice on technical issues. We thank JoAnne Stubbe for helpful discussions on experimental design and metabolic regulation. We thank Marcia Osburne for comments on the manuscript, and Jeffrey Palm for assistance in manuscript preparation. We acknowledge Matthew Henn and the Broad Institute for genome sequencing. This work was supported by the Gordon and Betty Moore Foundation, the Department of Energy (GTL), the National Science Foundation (C-MORE), and a NIH Training Grant to L.R.T.

Figure 3-4: (next page) Schematic model of *Prochlorococcus* metabolism in uninfected cells relative to infected cells. (a) Uninfected *Prochlorococcus* uses NADPH and ATP from the light reactions of photosynthesis to fix carbon dioxide in the Calvin cycle, producing net GAP during the day. This sugar is then oxidized in the PPP at night to R5P and NADPH, which are used for nucleotide biosynthesis for host genome replication. (b) Infected *Prochlorococcus* is influenced at several metabolic steps by cyanophage AMGs (red ovals). The light reactions, aided by AMGs for photosystem II (PSII), the plastocyanin pool (PC), photosystem I (PSI), and ferredoxin (Fd), lead to the production of NADPH and ATP, while AMGs for plastoquinol terminal oxidase (PTOX) and high-light inducible proteins dissipate excess light energy and stabilize the photosynthetic membrane. NADPH and ATP are not used to power carbon fixation because the Calvin cycle is blocked by phage-encoded CP12. This forces carbon flux through the PPP, which is aided by three AMGs for PPP enzymes. NADPH and ribose produced by the PPP, combined with NADPH and ATP produced by the light reactions, are used to power nucleotide biosynthesis, including ribonucleotide reductase (RNR). Reactions marked with asterisks (*) are not single steps.


a. Uninfected
Cell membrane
Cytoplasm
Thylakoid membrane
Lumen
CC/PPP
Calvin cycle
Pentose phosphate pathway
Light reactions
Nucleotide biosynthesis
NADP
NADPH
GAP
FBP
F6P
DHAP
G6P
BPG
ADP
ATP
SBP
E4P
S7P
6PGL
PGA
X5P
R5P
NDPs
Ru5P
6PG
RuBP
CO2
RNR
dNDPs
gDNA
B12 precursor + Co2+
B12
hv
H2O
O2 + H+
PTOX
PSII
PQH2
PQ
Cyt b6f
PCred
PCox
PSI
Fd
FNR
ATPase
HLIP
H+
b. Infected
talC
cp12
zwf
gnd
cobS
nrdA
nrdB
petF
petE
psa
psbA
psbD
hli

CHAPTER 4

Redox dynamics of *Prochlorococcus* under cyanophage infection

Luke R. Thompson, Qinglu Zeng, and Sallie W. Chisholm

(manuscript to be submitted)

Abstract

Viruses (cyanophage) infecting *Prochlorococcus* carry genes for the pentose phosphate pathway (PPP) enzyme transaldolase and the Calvin cycle inhibitor CP12, along with key proteins in the light reactions of photosynthesis. Both the PPP and the light reactions generate NADPH, whereas the Calvin cycle consumes NADPH. NADPH is a critical precursor for nucleotide biosynthesis, genes for which are also carried by cyanophage. To investigate what effect cyanophage infection, exploiting these metabolic pathways, might have on NADPH levels in *Prochlorococcus*, we measured the NADPH/NADP ratio and the related NADH/NAD ratio during infection of *Prochlorococcus* MED4 by cyanophage P-HM2. We also measured the phosphorylation state of the total NAD(P)(H) pool, which controls complex formation between CP12 and its binding targets in the Calvin cycle, phosphoribulokinase and glyceraldehyde-3-phosphate dehydrogenase. Infection was carried out in the light and in the dark to see how the shift to dark affects these ratios and how this, in turn, affects the infection process. Upon infection in the light, the NADPH/NADP ratio increased two-fold, while the NADH/NAD ratio was unaffected, consistent with increased activity of the light reactions and PPP and decreased Calvin cycle activity. In the dark, uninfected controls decreased in both NADPH/NADP and NADH/NAD, and phage infection increased NADPH/NADP only marginally, consistent with the importance of light for phage replication, possibly via NADPH production. The NADP(H)/NAD(H) ratio declined in phage-infected cultures in the light, conditions expected to favor CP12-PRK-

GAPDH complex formation and thereby suppression of the Calvin cycle and conservation of NADPH. The increase in NADPH/NADP in infected cells in the light may provide elevated reducing power to fuel nucleotide biosynthesis for phage genomic DNA replication.

Introduction

Cyanophage infecting *Prochlorococcus* and *Synechococcus* carry 'auxiliary metabolic genes' (AMGs), which encode enzymes and regulatory proteins thought to influence host metabolism during infection, leading to a more productive infection (Sullivan et al. 2005, Weigele et al. 2007, Millard et al. 2009). Cyanophage AMGs include genes involved in the light reactions of photosynthesis, such as photosystem II (Mann et al. 2003) and photosystem I (Sharon et al. 2009), but no genes for the Calvin cycle, although importantly they do carry genes for the Calvin cycle inhibitor CP12 (Chapter 3). Additional cyanophage AMGs are found for the pentose phosphate pathway (PPP) and nucleotide biosynthesis (Sullivan et al. 2005). These are the core metabolic pathways of cyanobacteria, involving production and consumption of the two primary energy currencies of cyanobacteria: adenosine triphosphate (ATP) and nicotinamide adenine dinucleotide phosphate (NADPH). A model of the role of AMGs in photosynthesis, carbon metabolism, and nucleotide biosynthesis was presented in Chapter 3 (p. 110). According to this model, AMGs for the light reactions of photosynthesis (Equation 4.1), inhibition of the Calvin cycle (Equation 4.2), and the PPP (Equation 4.3) are expressed during infection, with their proteins integrating into these host metabolic pathways, resulting in the net production of NADPH, ATP, and ribose (underlined in Equations 4.1 and 4.3) for nucleotide biosynthesis.

$$\text{Light} + H_2O + \text{NADP} + \text{ADP} \xrightarrow{\text{Light reactions}} O_2 + \underline{\text{NADPH}} + \underline{\text{ATP}} \tag{4.1}$$

$$CO_2 + \text{NADPH} + \text{ATP} \xrightarrow{\cancel{\text{Calvin cycle}}} \text{Glucose} + \text{NADP} + \text{ADP} \tag{4.2}$$

$$\text{Glucose} + \text{NADP} \xrightarrow{\text{PPP}} \underline{\text{Ribose}} + \underline{\text{NADPH}} + CO_2 \tag{4.3}$$

Nucleotide biosynthesis itself is putatively augmented by phage-encoded ribonucleotide reductase and other phage-encoded nucleotide biosynthesis enzymes.

AMGs for photosystem II, Calvin cycle inhibition, the PPP, and nucleotide biosynthesis have been shown to be transcribed during infection of *Synechococcus* by cyanophage

Table 4.1: Pyridine nucleotide abbreviations used in this chapter.

Abbreviation	Meaning
NAD	β-nicotinamide adenine dinucleotide (oxidized form)
NADH	β-nicotinamide adenine dinucleotide (reduced form)
NADP	β-nicotinamide adenine dinucleotide phosphate (oxidized form)
NADPH	β-nicotinamide adenine dinucleotide phosphate (reduced form)
NAD(H)	NAD + NADH
NADP(H)	NADP + NADPH
NAD(P)(H)	NAD + NADH + NADP + NADPH
NADH/NAD	Redox state of NAD(H) pool
NADPH/NADP	Redox state of NADP(H) pool
NADP(H)/NAD(H)	Phosphorylation state of total NAD(P)(H) pool

Syn9 (Chapter 3). Work in other cyanophages has shown that AMGs for photosystem II proteins are translated into protein in *Prochlorococcus* infected by cyanophage P-SSP7 (Lindell et al. 2005). Little is known, however, about the metabolic changes incurred under phage infection. There are no reports on changes in marine *Prochlorococcus* or *Synechococcus* metabolite levels upon infection by cyanophage. There are, however, reports on redox changes in freshwater cyanobacteria upon cyanophage infection. In *Synechococcus elongatus* PCC7942 (*Anacystis nidulans*), phage infection leads to a decrease in both the oxidized and reduced forms of NADPH, resulting in oligomerization and increased activity of glucose-6-phosphate dehydrogenase (G6PDH) (Cséke and Farkas 1979, Cséke et al. 1981). G6PDH produces NAPDH and is usually the rate-limiting step of the PPP (Luzzatto 1967). In *Nostoc muscorum*, phage infection causes the thioredoxin *m* pool to become more reduced and also increases G6PDH activity (Amla et al. 1987).

A unifying theme among many cyanophage AMGs is their relation to the redox cofactor NADPH (the reduced form of NADP; Table 4.1 and Figure C-1 on page 145). Several major metabolic pathways in cyanobacteria that affect the production or consumption of NADPH are represented by AMGs. Photosynthetic electron transport and the PPP produce NADPH, and CP12 inhibits phosphoribulokinase (PRK) and glyceraldehyde-3-phosphate dehydrogenase (GAPDH) in the Calvin cycle, thereby decreasing consumption of NADPH (Tamoi et al. 2005). The complex of CP12 with PRK and GAPDH is stabilized by NAD(H) and destabilized by NADP(H) (Wedel and Soll 1998). AMGs are also found for ribonucleotide reductase (RNR) (Sullivan et al. 2005), which uses NADPH as its terminal electron donor to reduce nucleotides (NDPs) to deoxynucleotides (dNDPs).

Synthesis of deoxynucleotides by cyanophage RNR is likely a critical step in the infection process. Every cyanophage infecting *Prochlorococcus* and *Synechococcus* carries a gene for RNR (Chapter 3), even those phages whose relatives from other hosts do not carry RNR genes (e.g., bacteriophage T7 of *E. coli*). DNA biosynthesis is critical for phage replication because the number of progeny phage an infecting phage can yield (i.e., the burst size) is proportional to the amount of DNA it can make. RNR, which requires reducing equivalents from NADPH, is the only pathway for conversion of ribonucleotides to deoxyribonucleotides and therefore is essential to all living cells (Jordan and Reichard 1998).

Given the dependence of phage DNA biosynthesis on NADPH, and the significant number of phage AMGs that either help produce NADPH or prevent its use in carbon fixation, we designed experiments to address the following questions: How does cyanophage infection affect the NADPH/NADP ratio of *Prochlorococcus*? What effect does light have on this ratio in relation to phage replication? This is a particularly salient question since photosynthesis is a major producer of NADPH and cyanophages express photosynthesis genes. We used as our model system *Prochlorococcus* MED4 and cyanophage P-HM2. More is known about *Prochlorococcus* MED4 with respect to the light–dark cycle and regulation of photosynthesis (Zinser et al. 2009) than any other strain of *Prochlorococcus*. Cyanophage P-HM2 was isolated on *Prochlorococcus* MED4 and carries the AMGs *talC* (transaldolase) and *cp12* (CP12)—the two most commonly found AMGs in the PPP and Calvin cycle (Chapter 3)—as well as several AMGs involved in the light reactions of photosynthesis (e.g., *psbA* and *hli*). We measured the redox state of the NADP(H) pool (NADPH/NADP ratio), the redox state of the NAD(H) pool (NADH/NAD ratio), and the phosphorylation state of the total NAD(P)(H) pool (NADP(H)/NAD(H) ratio) during infection of *Prochlorococcus* MED4 by cyanophage P-HM2 in the light and in the dark. We then interpreted these results in the context of replication dynamics of P-HM2 in the light and in the dark.

Materials & Methods

Infection of *Prochlorococcus* MED4 by cyanophage P-HM2

Axenic *Prochlorococcus* MED4 was maintained in Pro99 medium (Moore et al. 2007) made with filtered Sargasso seawater. Salts and metals for Pro99 medium were from Sigma-Aldrich (St. Louis, MO, USA). Cultures were grown in constant light of 90 μE m^{-2} s^{-1}

with cool white fluorescent lamps or constant dark. Temperature was maintained at 19–22°C. Log-phase *Prochlorococcus* MED4 (4×10^7 cells mL^{-1}) was infected with cyanophage P-HM2 (4×10^7 infective phage mL^{-1}), resulting in a multiplicity of infection (MOI) of 1. Cell concentration was determined by flow cytometry (Influx, Cytopeia-BD, Seattle, WA, USA), and phage concentration was determined by the most probable number (MPN) assay (Tillett 1987). For both light and dark experiments, which were conducted on separate days, two replicate infected cultures and two replicate uninfected control cultures of 2 L each were maintained. Uninfected controls were given spent medium instead of phage lysate. Both spent medium and phage lysate were filtered through 0.2-μm polycarbonate filters (Millipore, Billerica, MA, USA) prior to addition.

Following infection, cultures were placed in a dark incubator or returned to the light incubator. Samples were taken at regular intervals to be analyzed for RNA, genomic DNA (gDNA), and pyridine nucleotides. For phage and host gDNA quantification, 100-μL samples were filtered with 0.2-μm polycarbonate filters. The filtrate was diluted 1:1000 for extracellular phage gDNA quantification. For intracellular phage and host gDNA quantification, the filter was washed with three 1-mL volumes of preservation solution (10 mM Tris-HCl, 100 mM EDTA, 500 mM NaCl, pH 8.0) and flash frozen; the cells were subsequently resuspended in 650 μL 10-mM Tris-HCl (pH 8.0) by aggitation in a Mini-Beadbeater (BioSpec, Bartlesville, OK, USA), the supernatant heated to 95°C for 15 min, then diluted 1:100. For RNA and pyridine nucleotides samples, 200 mL axenic *Prochlorococcus* culture was harvested by centrifugation at 15,000×g for 10 min at 4°C, decanted, resuspended in approximately 1 mL supernatant, aliquoted equally into four tubes, and centrifuged again at 15,000×g for 5 min at 4 °C. Samples for RNA (two tubes) were flash frozen in liquid nitrogen and stored at –80°C. Samples for pyridine nucleotides (two tubes) were extracted fresh as described below.

Growth of *Prochlorococcus* MED4 on a light–dark cycle

Axenic *Prochlorococcus* MED4 was maintained in Pro99 medium made with filtered Sargasso seawater. Cultures were grown in a 'sunbox', a modified Percival Scientific (Boone, IA, USA) I-35LL plant growth chamber with a 24-h light–dark cycle consisting of 5 h of increasing light from 0–320 μE m^{-2} s^{-1}, 5 h of 320 μE m^{-2} s^{-1}, 4 h of decreasing light from 320–0 μE m^{-2} s^{-1}, and 10 h of dark (Figure 4-1). Temperature was maintained at 24 ± 0.2°C. Two bottles

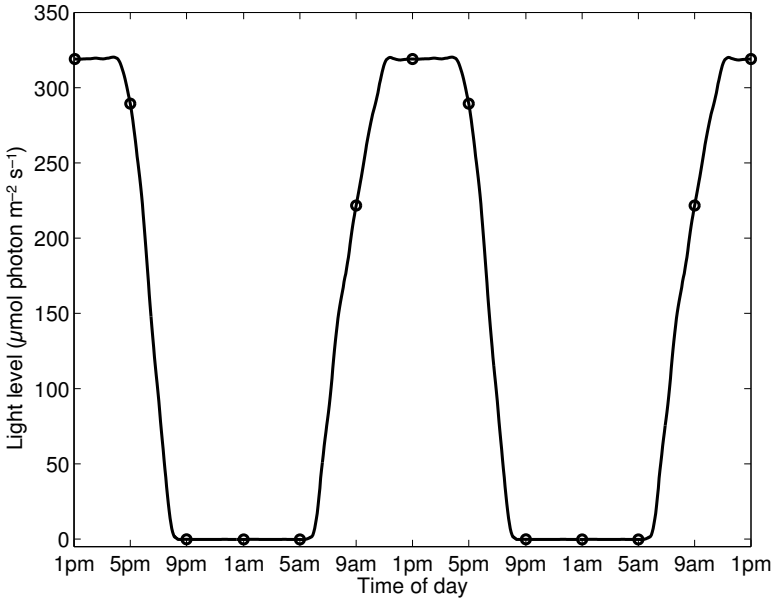


Figure 4-1: Light levels in the sunbox incubator over the diel cycle, with sampling points for pyridine nucleotide measurements marked with circles.

at a starting concentration of 3×10^7 cells mL^{-1}, as determined by flow cytometry, were sampled every 4 h for 48 h, starting at 1:00 pm (see Figure 4-1). For pyridine nucleotides, 120 mL axenic *Prochlorococcus* culture was harvested by centrifugation at 15,000$\times g$ for 10 min at 4°C, decanted, resuspended in approximately 1 mL supernatant, aliquoted equally into 2 tubes, and centrifuged again at 15,000$\times g$ for 5 min at 4°C. Pyridine nucleotides were extracted fresh as described below.

Quantitative PCR

qPCR primers were designed for *g20* (portal protein) from the genome of cyanophage P-HM2 (Sullivan et al. in press, Appendix G) and for *rnpB* (RNA component of ribonuclease P) from the genome of *Prochlorococcus* MED4 (GenBank) using Primer3 (Rozen and Skaletsky 2000) with a GC clamp of at least 2 bp, yielding products of 150–200 bp. Sequences are given in Table 4.2. Primers were tested using P-HM2 and MED4 gDNA and were shown to have specific and concentration-dependent amplification of target DNA.

Genomic DNA copies were quantified using the QuantiTect SYBR Green PCR Kit (QIAGEN, Valencia, CA, USA) with a LightCycler 480 Real-Time PCR System (Roche Diagnostics, Indianapolis, IN, USA). qPCR reactions contained 0.5 μM forward and reverse

Table 4.2: qPCR primers used in this study.

Gene	Forward primer	Reverse primer
Cyanophage P-HM2		
g20	5'-CGTAGAGAAGGTGGCAGAGG-3'	5'-GACCTTCCGATGTTAAATTGC-3'
Prochlorococcus MED4		
rnpB	5'-AAAGCAGGAGAGGCAATCG-3'	5'-TTAGGCGGTATGTTTCTGTGG-3'

primers and approximately 0.5 ng μL^{-1} cDNA. The amplification reaction consisted of an initial activation step of 15 min at 95°C, followed by 50 cycles of 15 s at 95°C (denaturation), 30 s at 56°C (annealing), and 30 s at 72°C (extension), followed by extension for 5 min at 72°C, followed by a melting curve from 50–90°C. Threshold cycle (C_T) of amplification was determined by the second derivative maximum method. Concentrations of phage and host gDNA over the time course were determined with standard curves of log(concentration of standard) versus C_T.

Measurement of pyridine nucleotides in *Prochlorococcus*

Pyridine nucleotides NAD, NADH, NADP, and NADPH were extracted and measured enzymatically as described previously (Maciejewska and Kacperska 1987, Leonardo et al. 1996, Tamoi et al. 2005). See page 147 for a detailed protocol and flowchart. NADH and NADPH standards were from Calbiochem (Gibbstown, NJ, USA), and all other reagents and enzymes were from Sigma-Aldrich (St. Louis, MO, USA). Pyridine nucleotides were extracted as follows. Fresh cell pellets were resuspended in 200 μL 100 mM HCl, 500 mM NaCl (for determination of NAD and NADP) or 200 μL 100 mM NaOH, 500 mM NaCl (for determination of NADPH and NADH). These resuspensions were then heated at 95°C for 5 min, centrifuged at 15,000×g for 5 min at 4°C, and the supernatants removed, flash frozen in liquid nitrogen, and stored at –80°C. NADH and NADPH standards of 20, 50, 100, 200, 500, and 1000 nM were prepared in 100 mM NaOH, 500 mM NaCl from stock solutions whose concentrations were determined by A_{340} (Cary 3 UV–visible spectrophotometer, Varian, Palo Alto, CA, USA).

All assays were carried out at 30°C in 200-μL reactions. Concentrated master solutions were made such that combining 180 μL master solution with 20 μL sample or unknown would yield the following final concentrations: 100 mM bicine (pH 8.0), 4 mM EDTA, 1.66

mM phenazine ethosulfate (PES), and 0.42 mM 3-(4,5-dimethyl-2-thiazolyl)-2,5-diphenyl-2H-tetrazolium bromide (MTT). For determination of NAD and NADH, master solutions also contained 10% ethanol and 0.2 U alcohol dehydrogenase (final concentrations). For determination of NADP and NADPH, master solutions also contained 5 mM glucose 6-phosphate and 0.2 U glucose-6-phosphate dehydrogenase (final concentrations). Assays were initiated by combining 180 μL master solution with 20 μL *Prochlorococcus* extract or 20 μL NADH or NADPH standard. All assays were performed in duplicate. Time-dependent increases in A_{550} were monitored using an Ultramark Microplate Reader (Bio-Rad, Hercules, CA, USA) for approximately 20 min. Absorbance data were smoothed using the robust loess method, and rates were calculated by linear regression analysis, implemented with MATLAB software (MathWorks, Natick, MA, USA). Standard curves were then used to calculate NAD, NADH, NADP, or NADPH concentrations, from which relevant ratios were calculated.

Results & Discussion

NADPH/NADP and NADH/NAD ratios

We first examined changes in NADPH/NADP and NADH/NAD ratios over the course of infection of cells infected in constant light. There was a dramatic increase in the NADPH/NADP ratio during infection relative to uninfected cells (Figure 4-2c), with the NADPH/NADP ratio increasing from ∼0.7 to ∼1.3 over the first 6 h and then remaining steady. There was no difference in the NADH/NAD ratio between infected and uninfected cells (Figure 4-2d), although there was a slight dip in this ratio for both treatments through 4 h followed by a slight increase up to 10 h.

The increase in NADPH/NADP ratio under infection in the light is consistent with increased flux through the PPP and/or decreased flux through the Calvin cycle. Since cyanophage genes for the PPP and for the Calvin cycle inhibitor CP12 are expressed during infection, we postulate that the action of these cyanophage gene products is partially responsible for this effect. An increase in the NADPH/NADP ratio is also consistent with increased photosystem activity, consistent with the light reaction genes being expressd during infection. Interestingly, during infection of freshwater *Synechococcus* by cyanophages not known to carry these genes for host metabolism, the NAPDH/NADP ratio was observed

to go down (Cséke and Farkas 1979, Cséke et al. 1981). Admittedly, it is difficult to assign the phenotype observed in infected *Prochlorococcus* to the activity of particular phage gene products. Without deleting individual AMGs from P-HM2, it is impossible to know what contribution, if any, a particular AMG is making to the increase in NADPH/NADP. It is also unclear why, if cyanophage need NADPH for DNA biosynthesis, the ratio does not go down or at least remain steady. One possible explanation is that an elevated steady-state concentration of NADPH is important for DNA biosynthesis. Modeling of metabolite fluxes will be instrumental in informing these possible scenarios.

As we noted above, the light reactions are heavily represented in cyanophage genomes, and they are thought to be important to the infection process. We know from previous experiments by our group and others (Sherman 1976, Lindell et al. 2005, Kao et al. 2005) that light is important for cyanophage infection. Lindell et al. (2005) showed that if cells are placed in the dark or if the light reactions are inhibited with the herbicide 3-(3,4-dichlorophenyl)-1,1-dimethylurea (DCMU), which blocks electron transfer from photosystem II to the plastoquinone pool, there is a significant decrease in the number of phage produced. Kao et al. (2005) also found that the number of phage progeny was correlated with the amount of light, and phage production was correlated with the light–dark cycle, with phage production increasing in the morning and cultures infected during the day producing more progeny. Part of the explanation for these observations could be that, in the dark, there is not enough NADPH for phage to reproduce. We suspected that the NAD(P)(H) pool would be more oxidized in the dark—as has been shown in freshwater cyanobacteria (Tamoi et al. 2005)—possibly hampering phage replication. We asked, how does the NADPH/NADP ratio change when cells are shifted from light to dark, and what effect does phage infection have on this?

The NADPH/NADP ratio decreased when infection was carried out in cultures moved to the dark (Figure 4-3c). In the uninfected control shifted to the dark, the NADPH/NADP ratio decreased steadily from ∼0.5 to ∼0.3 over 10 h. There was a similar decrease in the NADH/NAD ratio, though it was not a steady decrease (Figure 4-3c). In the uninfected control, the NADH/NAD ratio decreased from ∼1.4 to ∼0.8 over 10 h, with an initial decrease through 2 h, an increase at 4 h, and then a steady decrease until 10 h.

The decreases in NADPH/NADP and NADH/NAD ratios in the dark are indicative of the importance of light in maintaining *Prochlorococcus* in a reduced state. Without light,

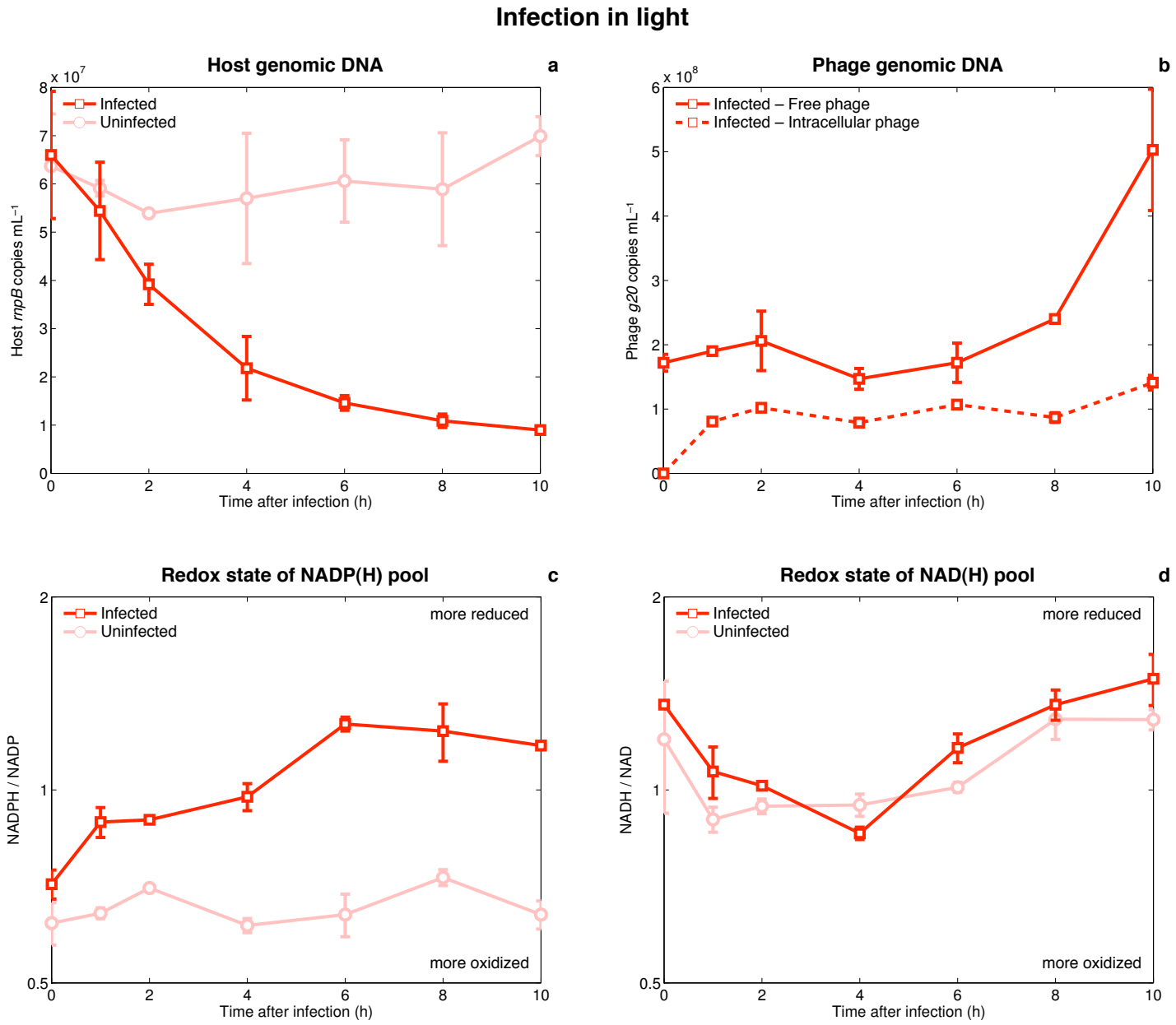


Figure 4-2: Infection dynamics, NADPH/NADP ratio, and NADH/NAD ratio during infection of *Prochlorococcus* MED4 by cyanophage P-HM2 in the light. (a) Degradation of host gDNA as determined by qPCR of *rnpB* copies. (b) Replication of phage gDNA as determined by qPCR of intracellular and extracellular *g20* copies. (c) Redox state of NADP(H) pool. (d) Redox state of NAD(H) pool.

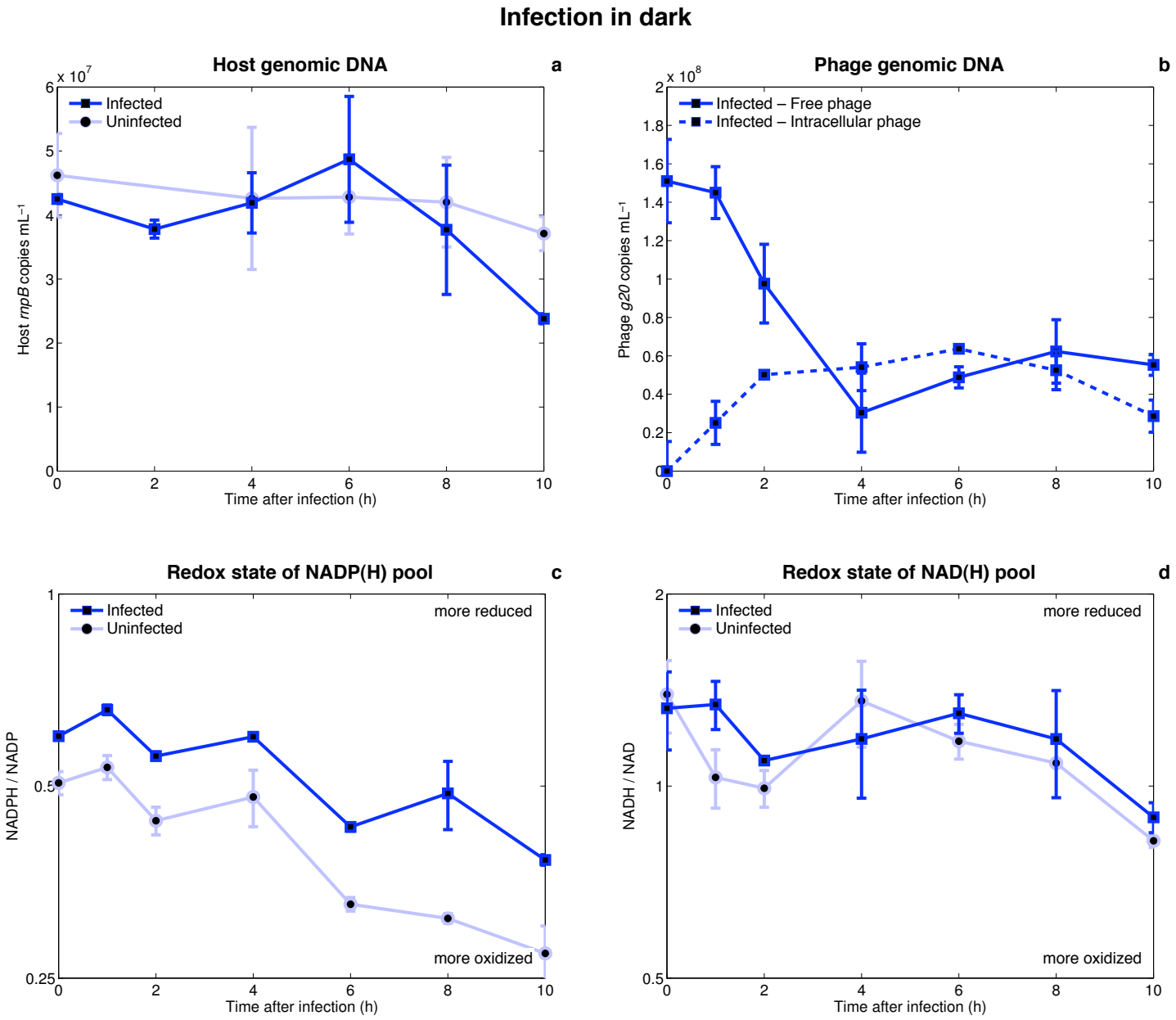


Figure 4-3: Infection dynamics, NADPH/NADP ratio, and NADH/NAD ratio during infection of *Prochlorococcus* MED4 by cyanophage P-HM2 in the dark. (a) Degradation of host gDNA as determined by qPCR of *rnpB* copies. (b) Replication of phage gDNA as determined by qPCR of intracellular and extracellular *g20* copies. (c) Redox state of NADP(H) pool. (d) Redox state of NAD(H) pool.

the photosynthetic electron transport chain is deprived of electrons to reduce NADP, while at the same time, the Calvin cycle and nucleotide biosynthesis may continue to consume NADPH. With a major source of NADPH turned off and significant sinks of NADPH still running, the NADPH/NADP ratio would be expected to drop. The decrease in the NADH/NAD ratio has a less obvious explanation, since the role of NADH in cyanobacteria is less clear. Cyanobacteria have an incomplete TCA cycle (Stanier and Cohen-Bazire 1977, Cooley et al. 2000), and it is unclear to what degree they oxidize glucose to NADH via the TCA cycle for ATP synthesis. The redox state of the NAD(H) pool, however, is linked to the redox state of the NADP(H) pool via pyridine nucleotide transhydrogenase, a membrane-associated protein complex that can transfer a hydride from NADH to NADP, coupled to proton translocation from the thylakoid lumen to the cytoplasm (Equation 4.4) (Prasad et al. 1999).

$$H^+_{\text{lumen}} + \text{NADH} + \text{NADP}^+ \rightleftharpoons H^+_{\text{cytoplasm}} + \text{NAD}^+ + \text{NADPH} \tag{4.4}$$

If these two pools (NAD(H) and NADP(H)) are linked, it makes sense that both the NADH/NAD and NADPH/NADP ratios would decrease following a shift from light to dark. NADH may serve as a reservoir of reducing power for the cell. As the NADP(H) pool becomes oxidized after photosynthesis shuts down, the reducing power in the NAD(H) pool could be transferred to the NADP(H) pool via the activity of transhydrogenase, keeping the NADP(H) pool from becoming overly oxidized. Hydride transfer from NADH to NADP via transhydrogenase is well documented in mitochondria (Pietro and Lang 1958, Prasad et al. 1999, Pedersen et al. 2008). This scenario is consistent with redox dynamics in *Prochlorococcus* over the diel cycle, in which the major changes observed were in the NADH/NAD ratio rather than the NADPH/NADP ratio (Appendix B), even though the major metabolic pathways in *Prochlorococcus* directly involve NADPH and NADP.

Phage infection did little to change the overall oxidizing trend in dark cultures, as the infected cultures saw decreases in the NADPH/NADP ratio (Figure 4-3c) and the NADH/NAD ratio (Figure 4-3d) similar to those seen in the uninfected controls. The NADPH/NADP ratio was also the only metric by which a difference could be discerned between infected and uninfected cultures, as the NADPH/NADP ratio in the infected cultures decreased slightly less than in the uninfected cultures. Thus, phage infection

in the dark has a similar (if less dramatic) effect as infection in the light: infection increases the NADPH/NADP ratio relative to the uninfected control, and it has no effect on the NADH/NAD ratio. Nevertheless, the absolute ratios of both NADPH/NADP and NADH/NAD decrease in the dark. The light-to-dark shift seems to trump the effect of phage infection on these ratios.

Phage replication

Given the oxidizing effect of darkness on *Prochlorococcus*, and the inability of these cells to photosynthesize in the dark, it is likely that cells in the dark would not yield as successful an infection as cells in the light, as has been shown previously for other cyanophages (Kao et al. 2005, Lindell et al. 2005). Indeed, infection in the dark was less robust than in the light. In the light, degradation of host gDNA under phage infection was rapid, dropping by 40% at 2 h and 65% at 4 h (Figure 4-2a), while in the dark no decrease in host gDNA relative to control was observed until 10 h (Figure 4-3a). In both light regimes, there was detectable intracellular phage gDNA after 1 h, and this did not increase significantly after 2 h (Figure 4-2b and Figure 4-3b). In the light, there was a rise in free phage at 8 h and 10 h (Figure 4-2b). In the dark, there was no increase in free phage over the span of the experiment (Figure 4-3b).

Rapid synthesis of phage gDNA and a strong phage burst in the light is consistent with a critical role for NADPH during infection. In the light, the increase in NADPH/NADP could provide the NADPH necessary to power DNA biosynthesis for phage replication. In the dark, the decrease in NADPH/NADP is correlated with a weak infection, in which there is delayed and minimal degradation of host gDNA, little detectable phage gDNA synthesis, and no detectable phage burst. Again, it is difficult to show a causal link between the NADPH/NADP ratio and successful infection. Yet it is clear that in P-HM2 infection of *Prochlorococcus* MED4, like in other cyanophage–host systems, phage reproduction is dependent on light.

Possible role of TalC and CP12

We propose that the action of TalC and CP12 encoded by P-HM2 is partly responsible for the increase in the NADPH/NADP ratio in infected cells in the light. If phage TalC and CP12 are indeed responsible for this effect, we would expect intracellular conditions to be

favorable for high TalC and CP12 activity. Specifically, formation of CP12-PRK-GAPDH complex is promoted by a decrease in NADP(H)/NAD(H), i.e., a decrease in the phosphorylation state of the total NAD(P)(H) pool (Tamoi et al. 2005). In the light, there was a small but significant decrease in the phosphorylation state of the total NAD(P)(H) pool in the infected treatment relative to the control (Figure 4-4a), with NADP(H)/NAD(H) at 10 h about one-third lower in the phage-infected cultures. In the dark, NADP(H)/NAD(H) decreased equally in control and infected cultures, but the decrease was more rapid than in the light (Figure 4-4b). To compare these ratios to the NADP(H)/NAD(H) ratio over the light–dark cycle, which is known to influence NADP(H)/NAD(H) and CP12 activity, dynamics of the NAD(P)(H) pool were measured over 48 hours of the diel cycle of *Prochlorococcus* MED4. These results are discussed in detail in Appendix B; here we present only the NADP(H)/NAD(H) ratio (Figure 4-4c).

If we compare the NADP(H)/NAD(H) ratio under infection in the light to that at night over the diel cycle (Figure 4-4c), when we expect CP12 to be active, we see that the final NADP(H)/NAD(H) ratio under infection is close to that at night but not quite as low (Figure 4-4). Further, the drop in NADP(H)/NAD(H) under infection is observed gradually over the experiment, whereas the most dramatic increase in NADPH/NADP (which is the presumed effect of CP12 activity, i.e., shutdown of the Calvin cycle) is observed from 0–1 h. One possible explanation for this disconnect is that the increase in NADPH/NADP later in infection is via CP12 activity, whereas the increase over the first few hours is due to other factors. Details of timing aside, the gross effect of cyanophage infection of *Prochlorococcus*—a decrease in the phosphorylation state of the total NAD(P)(H) pool—mirrors what was observed in cyanophage infection of freshwater *Synechococcus* (Cséke and Farkas 1979, Cséke et al. 1981).

Model of host metabolism under cyanophage infection

As noted above, the increase in the NADPH/NADP ratio during infection is consistent with increased production of NADPH as a result of an AMG-powered photosynthetic light reaction and PPP and an AMG-inhibited Calvin cycle. The nearly two-fold increase in NADPH/NADP in infected cultures in the light indicates that cyanophage infection dramatically alters the redox poise of *Prochlorococcus*. The activity of phage TalC and CP12 could help account for this. Properties of the NAD(P)(H) pool appear to be favorable

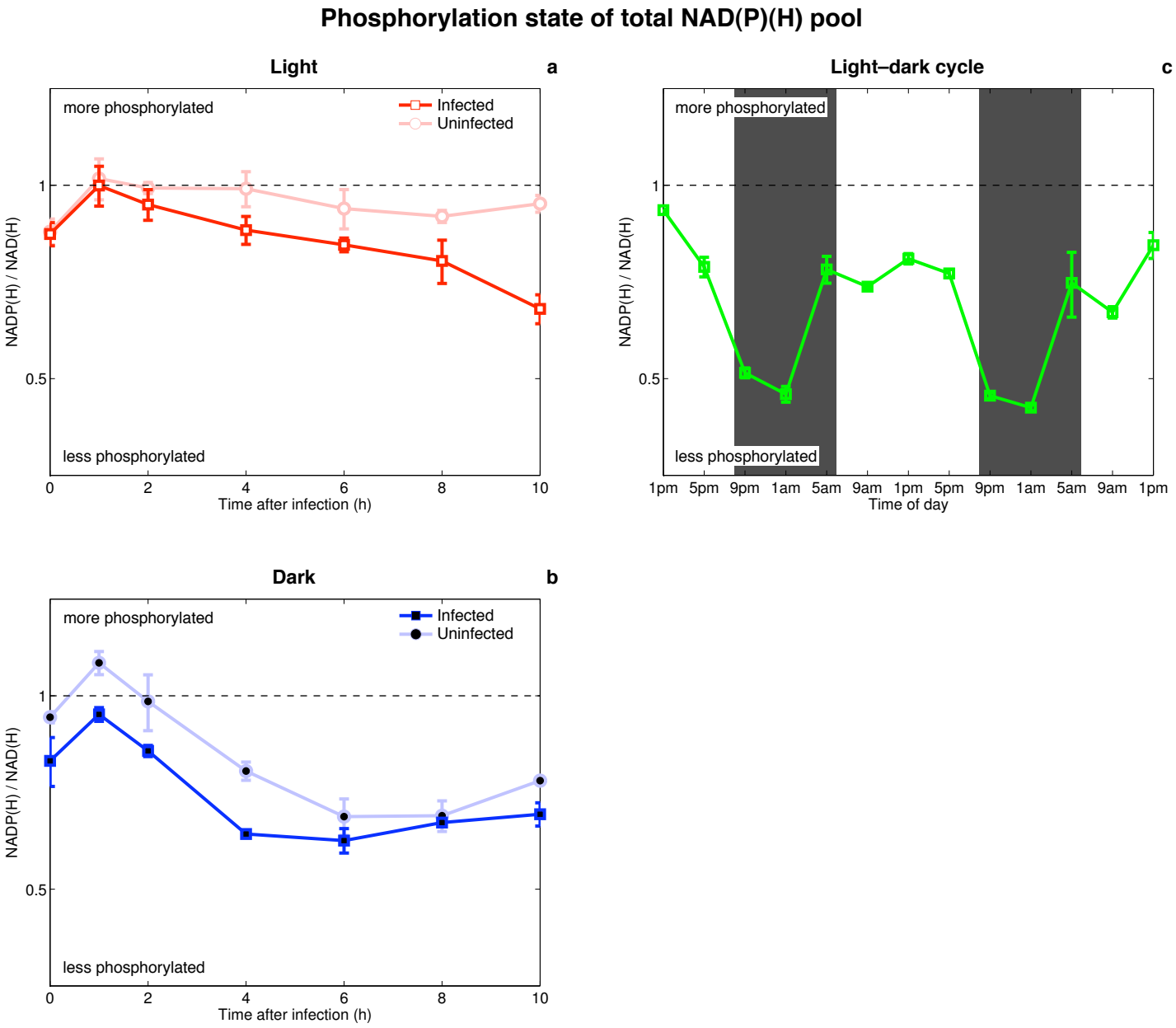


Figure 4-4: NADP(H)/NAD(H) ratio of *Prochlorococcus* MED4 under infection by cyanophage P-HM2 (a) in the light, (b) in the dark, and (c) over the diel cycle. Light levels over the diel cycle are given in Figure 4-1.

for CP12 binding and therefore activity. Specifically, the decrease in NADP(H)/NAD(H) ratio in infected cells would help stimulate CP12-PRK-GAPDH complex formation (Tamoi et al. 2005). NADPH produced by the TalC-enhanced PPP and the photosynthetic electron transport chain should not be consumed by a CP12-inhibited Calvin cycle. Rather, NADPH could accumulate to power phage nucleotide biosynthesis using phage-encoded ribonucleotide reductase and host-provided machinery. Metabolic flux along these lines may be critical for phage genome replication.

A model summarizing the hypothesized fluxes under the four regimes tested here (light or dark, infected or control) is shown in Figure 4-5. In uninfected cells in the light (Figure 4-5a), NADPH and ATP from the light reactions feeds directly into the Calvin cycle for storage in glucose, and DNA can be synthesized using NADPH and ribose from either the Calvin cycle or the PPP. Under infection in the light (Figure 4-5b), NADPH and ATP from the light reactions (bolstered by phage proteins) are not consumed by the Calvin cycle (inhibited by CP12) but rather feed into DNA biosynthesis, along with NADPH and ribose from the PPP (bolstered by TalC), and there is a net increase in NADPH/NADP. In uninfected cells in the dark (Figure 4-5c), the light reactions are unable to function, leading to a decrease in NADPH/NADP (and NADH/NAD) and shutdown of the Calvin cycle, and less DNA can be synthesized with this limited NADPH and ribose. Under infection in the dark (Figure 4-5d), the PPP is bolstered by TalC, with a moderate increase in NADPH/NADP relative to uninfected cells in the dark, but it is not enough to prevent the overall oxidizing effect of darkness and decreased capacity for nucleotide biosynthesis.

The pathways implicated by cyanophage genes and the NAD(P)(H) dynamics under infection suggest that redox is a critical variable in the infection of cyanobacteria by viruses. Turning to other photosynthetic organisms, there is also a redox link, but in many of these cases infection leads to a more oxidized cellular state, the opposite of what we observed. Viruses of green plants and algae are known to cause photosystem damage, formation of reactive oxygen species, and lipid peroxidation (Rahoutei et al. 2000, Arias et al. 2003, Evans et al. 2006). Freshwater cyanophages are known to shift freshwater *Synechococcus* to a more oxidizing state, and this leads to increased glucose-6-phosphate dehydrogenase activity (Balogh et al. 1979, Cséke et al. 1981) and likely increased flux through the PPP. Even in the *Prochlorococcus*–cyanophage system, damage to the *Prochlorococcus* photosynthetic apparatus by cyanophages and the resulting oxidative stress has been proposed as a major

reason why cyanophage carry *psbA* and *hli* genes (Lindell et al. 2004).

How can we resolve this apparent paradox? First, we need to distinguish 'oxidized state' from 'oxidative stress'. Photosynthesis generates NADPH (a reduced species), but it also generates reactive oxygen species (ROS, i.e., oxidized species), so the two can and often do coexist. It is possible, therefore, that part of the advantage of generating NADPH for cyanophage is in alleviating oxidative stress. Indeed, many cyanophages encode a plastoquinol terminal oxidase (PTOX), which is essentially a safety valve to consume electrons from an over-worked photosynthetic electron transport chain, preventing the generation of ROS (Mackey et al. 2008, Bailey et al. 2008, Bagby 2009, Latifi et al. 2009). Second, it is possible that the proposed metabolic hijacking strategy of cyanophage is a novel adaptation to the stresses imposed on the host by infection. To our knowledge, freshwater cyanophages and eukaryotic viruses do not encode transaldolase or CP12. Perhaps in response to the oxidative stress induced by infection, marine cyanophages have acquired these components and evolved a strategy to take over host carbon metabolism to harvest reducing equivalents from host glucose stores. The proper testing of this model awaits a genetic system in *Prochlorococcus* and cyanophage to knock out *talC* and *cp12*. The data presented here, however, serve to further highlight the importance of redox poise in the cyanobacteria–cyanophage symbiosis.

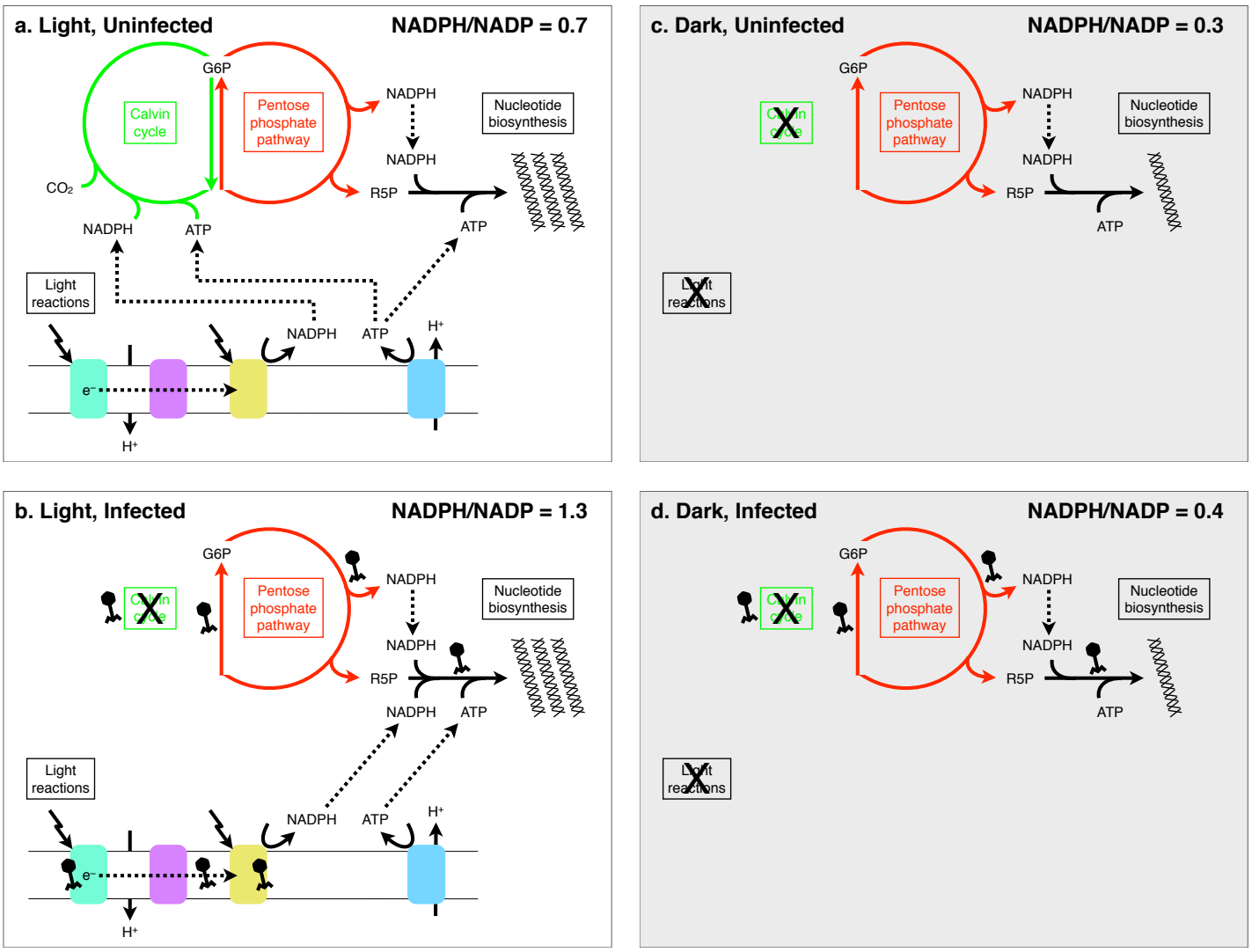


Figure 4-5: Model of *Prochlorococcus* metabolism under cyanophage infection in the light and in the dark. Phage icons denote metabolic steps proposed to be augmented (or inhibited, in the case of CP12) by the action of phage gene products. (a) Uninfected cells in the light: both the light reactions and the PPP produce energy, while the Calvin cycle stores some of that energy. (b) Infected cells in the light: the Calvin cycle is turned off by phage CP12, while phage-boosted light reactions and PPP produce energy (increased NADPH/NADP) for phage DNA replication. (c) Uninfected cells in the dark: the light reactions and Calvin cycle no longer produce energy (decreased NADPH/NADP). (d) Infected cells in the dark: phage boost the PPP, but without the light reactions there is not enough energy (decreased NADPH/NADP) to fuel sufficient phage DNA replication.

Acknowledgments

We thank JoAnne Stubbe, Alexa Price-Whelan, Dianne Newman, Alex Bradley, and Chris Marx for discussions on pyridine nucleotide assays. We thank Sara Roggensack for assistance in maintenance of cultures. This work was supported by the Gordon and Betty Moore Foundation, the Department of Energy (GTL), the National Science Foundation (C-MORE), and a NIH Training Grant to L.R.T.

CHAPTER 5

Conclusions and future directions

Conclusions

This thesis has highlighted the role of host-like metabolic genes, termed 'auxiliary metabolic genes' (AMGs), carried by marine cyanophages. The results and analyses presented here argue that the pentose phosphate pathway (PPP)—augmented by AMGs for this pathway—plays an important role in the replication of cyanophages in *Prochlorococcus* and *Synechococcus* hosts. The principal findings include the following:

- A phage-encoded transaldolase is functional and may therefore substitute for the transaldolase in the host. The phage transaldolase does not, however, exhibit an obvious kinetic advantage over the host transaldolase. Rather, phage use of its own transaldolase may derive from selection on phage genome size or host regulation of protein levels over the light–dark cycle.
- Prevalence patterns of PPP genes in phage genomes from cultured isolates and from the environment suggest a hierarchy of selective pressures for the maintenance of these genes over evolutionary time and possible metabolic bottlenecks during infection.
- Coordinated expression of phage PPP genes during infection with phage genes for the light reactions of photosynthesis and inhibition of the Calvin cycle suggests that cyanophage may disrupt the usual regulation of these pathways in the host. Relative to host replication, phage replication places an increased demand on DNA biosynthesis, suggesting that a major role of AMG-encoded proteins is to provide nucleotides for phage genome replication.
- The NADPH/NADP ratio of *Prochlorococcus* increases under infection, indicating that the modulation of host metabolic pathways during infection has demonstrable effects on the redox state of the host.

Future directions

Investigation of the metabolism of phage-infected cyanobacteria from multiple perspectives—phage gene content and gene expression, in vitro protein function, host metabolite levels, and phage replication dynamics—reveals the systemic reach of phage infection on host metabolism. The model of host metabolism under infection presented here attempts to integrate the functions of phage AMGs from multiple pathways, all coordinated for phage replication. This model, while useful, is limited in scope and quantitative rigor. The current model does not account for the complete host metabolism (with or without infection), is not quantitative, and is not informed by key parameters such as phage burst size, complete mRNA/protein/metabolite levels, and metabolic control coefficients. Nevertheless, with the small genomes of *Prochlorococcus* and cyanophages and the advent of new tools for measuring cellular dynamics, the building of a comprehensive model of host–phage metabolism is within reach.

Derivation of a comprehensive model of cyanophage infection should proceed along multiple avenues. First, the functions of phage-encoded proteins central to our nascent model should be investigated. Biochemical investigations should be focused on demonstrating the inhibitory effect of phage CP12 on host Calvin cycle enzymes and on the source of the diferric cluster and reducing equivalents for phage class Ia RNR. Second, parameters of phage infection must be measured in culture. These include the timing of phage adsorption, DNA translocation, DNA replication, and cell lysis, as well as phage burst size. Third, transcriptome, proteome, and metabolome dynamics of host–phage metabolism over the entire infection are required to inform the model. These measurements are possible with current technologies. Finally, all experiments in culture should be conducted over a light–dark cycle, which is the natural light regime of cyanobacteria and which places specific constraints on host metabolism, particularly with respect to energy and carbon metabolism. With these data, a model of *Prochlorococcus* metabolism over the light–dark cycle and under phage infection could be generated, tested, and refined. A greater understanding of the dynamics of cyanophage infection will inform the selective pressures that influence cyanobacterial and cyanophage populations, in turn informing their ocean-wide biogeochemical effects.

APPENDIX A

Supplementary material for Chapter 3

Luke R. Thompson, Qinglu Zeng, Libusha Kelly, Katherine H. Huang,
Maureen L. Coleman, and Sallie W. Chisholm

Materials & Methods

CP12 hydrophobicity plots

Hydrophobicity plots were used to compare hydrophobicity patterns of hypothetical cyanophage CP12 sequences and known CP12 sequences from plants, algae, and cyanobacteria, using the method of Kyte and Doolittle (1982) with a window size of 11 residues. Prior to plotting hydrophobicity, sequences were aligned using MUSCLE (Edgar 2004) and then trimmed to remove all gapped positions.

Genome alignment and phylogenetics

Genomic context of 6-phosphogluconate dehydrogenase (*gnd*) and glucose-6-phosphate dehydrogenase (*zwf*) was investigated using MicrobesOnline (Alm et al. 2005). Cyanophage *gnd* was used to anchor a genome alignment of seven cyanophage genomes and other closely related sequences, which had a conserved gene order (*gnd* followed by *zwf*). Two of these genomes, *Anabaena variabilis* ATCC29413 and *Agrobacterium tumefaciens* C58, were used in the final alignment; in addition to syntenic cyanophage-like *gnd* and *zwf*, both genomes have one additional copy each of *gnd* and *zwf*. By comparing the gene and species trees generated by MicrobesOnline, it was possible to determine whether particular *gnd* or *zwf* variants were 'native' (found in every or nearly every genome within a taxonomic group) or 'non-native' (found sporadically within a taxonomic group). Also included in the genome alignment were one *Synechococcus* and two *Prochlorococcus* genomes, whose *gnd* and *zwf* are

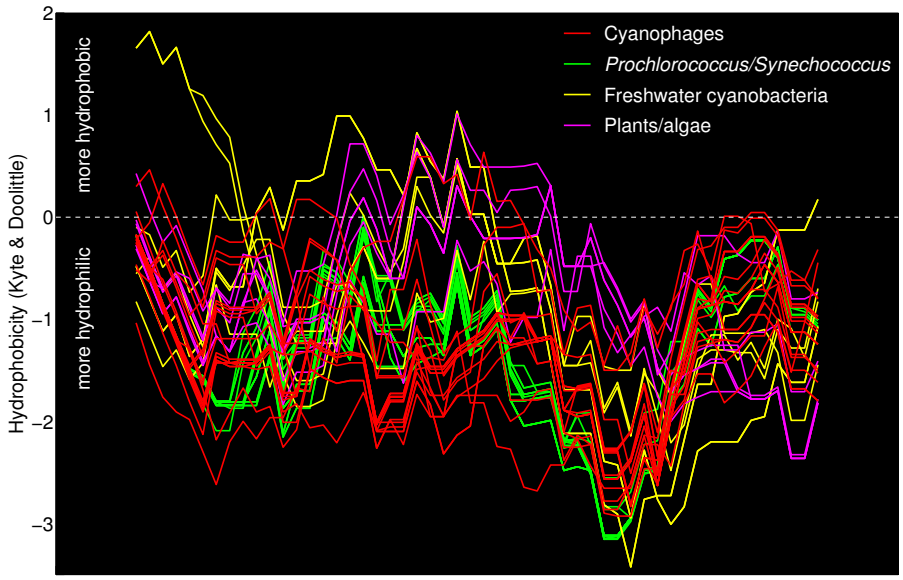
not syntenic.

Gnd and Zwf protein sequences were aligned separately using MUSCLE (Edgar 2004). Positions with gaps were removed if at least 50% of sequences had gaps at that position. The two multiple sequence alignments were then concatenated. Phylogenetic trees were built using the maximum likelihood algorithm implemented by PhyML (Guindon and Gascuel 2003), and statistical tests of branches were done using aLRT (approximate likelihood-ratio test) parametric statistics with Chi2-based parametric branch supports (Anisimova and Gascuel 2006). The following parameters were used: phyml_alrt alignment.phy 1 i 1 -2 JTT 0.0 4 e BIONJ y y.

Results & Discussion

Cyanophages encode the Calvin cycle inhibitor CP12

The gene for *cp12* was present in cyanophage genomes published as early as 2005 (Sullivan et al. 2005), yet its presence has gone unrecognized until now. Why was *cp12* overlooked in previous studies of cyanophage genomes (Sullivan et al. 2005, Weigele et al. 2007, Millard et al. 2009)? The 8-kDa CP12 protein is intrinsically unstructured, with little sequence conservation across plants and cyanobacteria (as low as 10% pairwise amino acid identity). As a result, CP12 was assigned a function in *Prochlorococcus* only recently, by manual analysis (Zinser et al. 2009). There is, nevertheless, a universal C-terminal motif, CxxxPxxxxC, that is found in all plants, algae, *Prochlorococcus*, *Synechococcus*, and cyanophage CP12 sequences, as well as a conserved pattern of hydrophilic amino acids (Figure A-1), a signature feature of this naturally unfolded protein. The C-terminal disulfide bond formed by the two cysteines in this motif is thought to facilitate binding to glyceraldehyde-3-phosphate dehydrogenase (Lebreton et al. 2006). Two N-terminal cysteines also form a disulfide bond in CP12 from plants and algae, facilitating binding to phosphoribulokinase (Pohlmeyer et al. 1996). Most cyanobacterial and cyanophage CP12 sequences lack the N-terminal cysteine (Figure A-1), with the exception of *Synechocystis*. Cyanobacterial CP12 (*Synechococcus* PCC7942) is nevertheless able to bind phosphoribulokinase (Tamoi et al. 2005), and we might expect cyanophage CP12 to have this function as well. The hydrophobicity pattern in CP12 is observed despite low sequence identity among this group.




```
Syn9        SIEKHIQADKEILQNPTQMRRHIESELDQLERYHESHPEHHDPLELYCEDEPGAPECKIHDN
P-SSM4      TIEQHIQYDKDLIENPMAARRHAKAELHELEEYAEHHKAHHDPLEIFCDLHPDEPECLIYDD
P-SS2       MRQAYFEEDSRQPSSTPPFIRASWSLLPRLGFNTDQATKQSLPFQSFCDMNPDSVRCREFDV
WH8109-1    SIDEHIQKDQSEIEAAKAKVRHLTDELKSLEEYKEHHPDSHDPLELHCEANPDADECRVYDD
WH8109-2    TIDEHIQKDQEEFLKALGKVRHLTEELQWLLDHKKEFPDPHDPLELFCEQNPDEPECLVYDD
MED4        SIDEHIQKDQSEIASAKPKVRHLTEELKDLEGYKDHHPEKHDPLELFCDANPDEPECLVYDD
PCC7942     KIEQARQEAHAISEEKGPDAAAAWDAVEELQAEAAHQRQSETEFGDYCSENPDAAECLIYDD
PCC6803     KIEQELANARQVCSTDEAECAAAWDAVEELEAEAAHQRQHPTQLEKFCDENPDAAECRIYDD
Chlamy      KVQDAVKEAEDACAKGTADCAVAWDTVEELSAAVSHKKDLTDPLEAFCKDAPDADECRVYED
Arab-2      VVEKSIKEAQETCAGDPGECVAAWDEVEELSAAASHARDGSDPLEEYCKDNPETNECRTYDN
              :                         :  *             :  .*.  *   .*  .:
```

Figure A-1: Hydrophobicity plot and ungapped alignment of CP12 from plants, algae, cyanobacteria, and cyanophages. Representative sequences from the alignment are shown. Abbreviations: WH8109, *Synechococcus* WH8109; MED4, *Prochlorococcus* MED4; PCC7942, *Synechococcus elongatus* PCC7942; PCC6803, *Synechocystis* PCC6803; Chlamy, *Chlamydomonas reinhardtii*; Arab, *Arabidopsis thaliana*. "-1" or "-2" following the sequence name denotes multiple copies in the genome.

Phage gene cassette for pentose phosphate pathway is found sporadically in plasmids and chromosomes

Gene-gain and gene-loss events could also help explain the distribution of PPP genes in cyanophages, independent of genome size. *Synechococcus* T4-like cyanophage genomes (174.4–232.9 kbp) and *Prochlorococcus* T4-like cyanophage genomes (176.4–252.4 kbp) are not significantly different in size, so genome size cannot explain why *Synechococcus* T4-like phages have *gnd* and *zwf* but *Prochlorococcus* T4-like phages do not. Interestingly, other phylogenetic and comparative genomic analyses on T4-like cyanophage genomes have also failed to find significant differences between *Prochlorococcus* and *Synechococcus* T4-like phages (Sullivan et al. in press, Appendix G), so the presence of *gnd* and *zwf* in only *Synechococcus* T4-like phages is notable. Noting that *gnd* and *zwf* are always adjacent in the phage genomes, we hypothesized that they might have been acquired together in a single event prior to the divergence of *Synechococcus* T4-like cyanophages, and this could help explain why only those phages have *gnd* and *zwf*. If *gnd* and *zwf* were acquired together in a single event by an ancestral *Synechococcus* T4-like phage, we would not expect to see them in other cyanophages. Supporting this hypothesis, both Sullivan et al. (in press, Appendix G) and Millard et al. (2009) have noted that *gnd* and *zwf* reside in a putative mobile gene cassette subject to horizontal gene transfer (HGT); this mobile gene cassette also contains *petE*, PTOX, and an endonuclease. Sullivan et al. further describe how *zwf* appears in a highly degraded form in cyanophages S-ShM2 and Syn1, yet still adjacent to *gnd*. We went a step further to see if the *gnd–zwf* gene arrangement could be found in other genomes.

To examine the bacterial origin of phage *gnd* and *zwf*, we searched for these genes using BLAST against all sequenced prokaryotic genomes. Notably, these two genes were found together in several plasmids. Further, when gene trees for phage-like *gnd* or *zwf* were compared to corresponding species trees for those genomes (see methods), the gene trees did not match the species trees. This came in contrast to host-like (cyanobacterial) *gnd* and *zwf*, whose gene trees closely matched the species tree for cyanobacteria. In other words, the phage-type genes were found only sporadically across cyanobacteria and other groups, in mobile genetic elements as well as chromosomes, implying HGT ('non-native' gene copies), whereas the host-type genes were found in all cyanobacterial genomes and only in chromosomes, implying vertical descent ('native' gene copies). Phylogenetic analy-

sis of concatenated Gnd and Zwf protein sequences from cyanophages, *Prochlorococcus* and *Synechococcus* (which have chromosome-encoded *gnd* and *zwf* only), and *Anabaena variabilis* ATCC29413 and *Agrobacterium vitis* S4 (which have chromosome- and plasmid-encoded *gnd* and *zwf*), showed that the phage copies were indeed more similar to the plasmid-encoded, 'non-native' copies than to the chromosome-encoded, 'native' copies (Figure A-2). Thus, cyanophage *gnd–zwf* appears to have been horizontally transferred and incorporated sporadically into bacterial genomes, including other mobile genetic elements such as plasmids. This could help explain why *gnd* and *zwf* are a unique feature of T4-like cyanophages isolated on *Synechococcus*, and it once again implicates cyanophages as important agents of HGT.

Cyanophage *gnd* and *zwf* are an example of how bacteriophage and bacteria can influence each other over evolutionary time. Particularly interesting here is the instance in *Anabaena* and other genomes of two copies each of *gnd* and *zwf*. How did these genomes get two copies? The most straightforward and naive answer would be gene duplication, but upon further examination—by phylogenetics and by genomics—it appears that the bacterial genomes most likely acquired the second copies (with both genes at the same time) via HGT: the gene phylogeny does not follow the species phylogeny, and the second copies reside in a mobile genetic element. All evidence supports that this second copy (*gnd–zwf* as a unit) is quite 'mobile' and is shared across phages, plasmids, and genomic chromosomes. It remains unclear whether cyanophage got this mobile element from these particular bacterial genomes or close relatives, or these bacteria got the mobile element from cyanophage, or whether there has been genetic exchange back and forth.

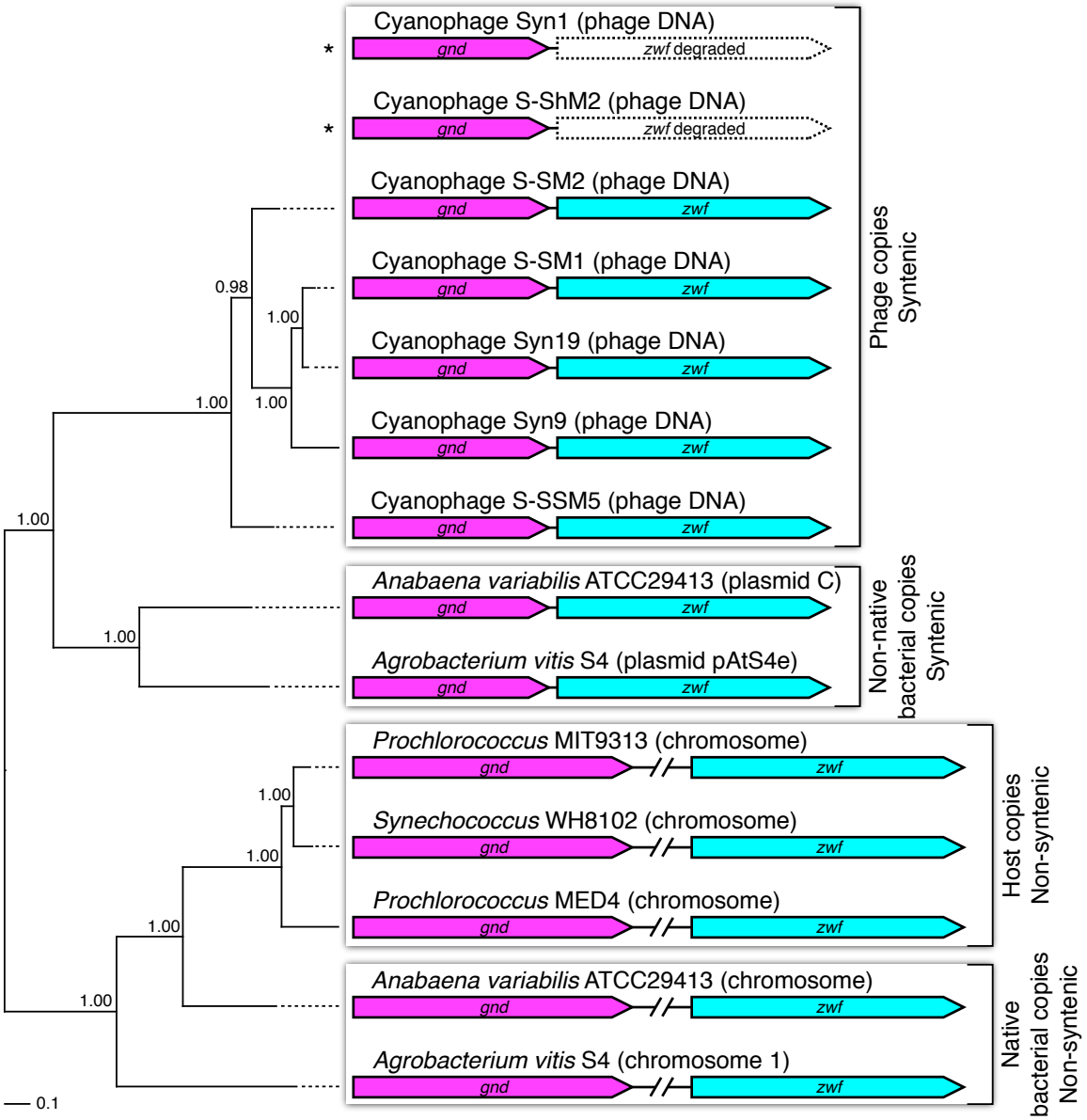


Figure A-2: Syntenic orthologs of cyanophage *gnd* (6-phosphogluconate dehydrogenase) and *zwf* (glucose-6-phosphate dehydrogenase). A maximum likelihood tree of concatenated Gnd and Zwf protein sequences is shown alongside the genomic position of *gnd* and *zwf* genes. The tree is midpoint-rooted and shown with Chi2-based parametric branch supports (see methods).

APPENDIX B

Supplementary material for Chapter 4

Luke R. Thompson, Qinglu Zeng, and Sallie W. Chisholm

Materials & Methods

Materials and methods are described in Chapter 4: growth of *Prochlorococcus* is described on page 117, and extraction and measurement of NAD, NADH, NADP, and NADPH are described on page 119.

Results & Discussion

NADPH/NADP and NADH/NAD ratios over the diel cycle

We measured levels of NAD, NADH, NADP, and NADPH in *Prochlorococcus* MED4 over the diel cycle. Based on studies of the diel cycle in cyanobacteria (Toepel et al. 2008, Stöckel et al. 2008, Zinser et al. 2009) and investigations of NAD(P)(H) levels over limited light–dark regimes (Tamoi et al. 2005), we expected to see diel variations in redox state (NADH/NAD and NADPH/NADP ratios) and phosphorylation state (NADP(H)/NAD(H) ratio) of the total NAD(P)(H) pool. Namely, we expected to see the NAD(P)(H) pool more reduced during the day, more oxidized at night, more phosphorylated during the day, and less phosphorylated at night.

We sampled *Prochlorococcus* MED4 grown in a 'sunbox' incubator at six times per day for 48 hours, for a total of 13 measurements (Figure 4-1). We found a dramatic diel variation in the NADH/NAD ratio (Figure B-1a), ranging from ∼2.0 just after sunset to ∼0.5 just before sunrise. The NADPH/NADP ratio (Figure B-1b), however, held constant at ∼0.5 over the diel cycle. Variation in the redox state of the total NAD(P)(H) pool (Figure B-1c), therefore, is due almost exclusively to changes in the NADH/NAD ratio.

The phosphorylation state of the total NAD(P)(H) pool (Figure 4-4c) exhibited a less dramatic but significant diel variation, with NADP(H)/NAD(H) ranging from ∼0.8 during the day to ∼0.4 at night.

The diel pattern in phosphorylation state of the NAD(P)(H) pool is consistent with regulation of CP12 by the NADP(H)/NAD(H) ratio. Recall that CP12 binding and inactivation of PRK and GAPDH in the Calvin cycle is favored by low NADP(H)/NAD(H) and disfavored by high NADP(H)/NAD(H). During the day, when CP12 should be inactive, allowing flow through the Calvin cycle, the NADP(H)/NAD(H) ratio is high. During the night, when CP12 should be active, inhibiting the Calvin cycle to promote the PPP, the NADP(H)/NAD(H) ratio is low.

The increase in reduction of the total NAD(P)(H) pool during the day was expected, but it was surprising that this change was observed in the NAD(H) pool and not in the NADP(H) pool. Daytime photosynthetic electron transport, specifically activity of photosystem I and the ferredoxin–NADP reductase, reduces NADP to NADPH (Blankenship 2002). Yet interestingly, we do not see an increase in NADPH/NADP during the day; rather, the increase is in NADH/NAD. One possible explanation is that the NAD(H) pool serves as a reservoir for electrons extracted from the photosynthetic electron transport chain by NADP. By passing electrons off to NAD, the NADP(H) pool can remain relatively oxidized and continue to be a favorable acceptor for photosynthetic electron flow.

A mechanism for how this might occur involves the pyridine nucleotide transhydrogenase (TH). TH is a thylakoid-associated complex that can reduce NADP using electrons from NADH and can also perform the reverse reaction (Pietro and Lang 1958). As shown in Equation B.1, TH reversibly transfers a hydride from NADH to NADP, forming NAD and NADPH, coupled to the proton-motive force.

$$\mathrm{H^+_{lumen}} + \mathrm{NADH} + \mathrm{NADP^+} \rightleftharpoons \mathrm{H^+_{cytoplasm}} + \mathrm{NAD^+} + \mathrm{NADPH} \tag{B.1}$$

Depending on the state of the proton gradient (Δp) and the concentrations of pyridine nucleotides, the reaction can run in either direction. During the day, the reaction could run in 'reverse' to produce NADH from NADPH and generate increased Δp; at night, this Δp could be used run TH 'forward' to power conversion of NADH back to NADPH for reductive biosynthesis. TH is maximally expressed at night (mRNA max: 4–6 pm) (Zinser

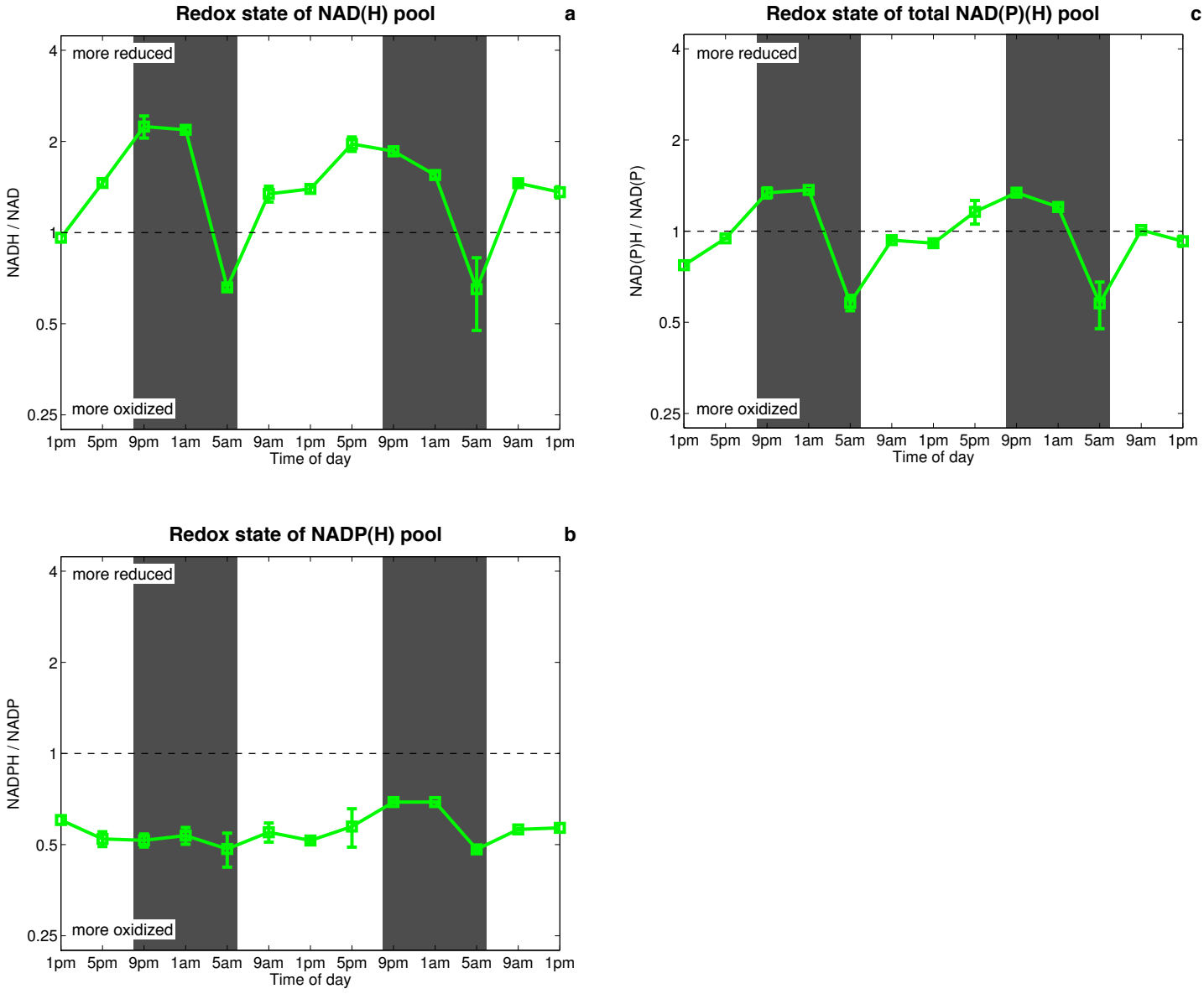


Figure B-1: NADH/NAD ratio, NADPH/NADP ratio, and NAD(P)H/NAD(P) ratio of *Prochlorococcus* MED4 over the diel cycle. (a) Redox state of NAD(H) pool. (b) Redox state of NADP(H) pool. (c) Redox state of total NAD(P)(H) pool. Light levels over the diel cycle are given in Figure 4-1.

et al. 2009), consistent with its functioning to transfer reducing equivalents from NADH to NADP.

This scenario is supported by NAD(P)(H) dynamics following the shift from constant light to constant dark (Figure 4-3). In that experiment, the NADPH/NADP ratio decreases steadily from 1 h after the shift to dark through the end of the experiment, while the NADH/NAD ratio fluctuates until 6 h after the shift to dark, when it also begins to decline steadily. This could be due to a two-step process, in which NADPH is first oxidized by pathways carrying out reductive biosynthesis, and then NADH is oxidized to replace NADPH via TH.

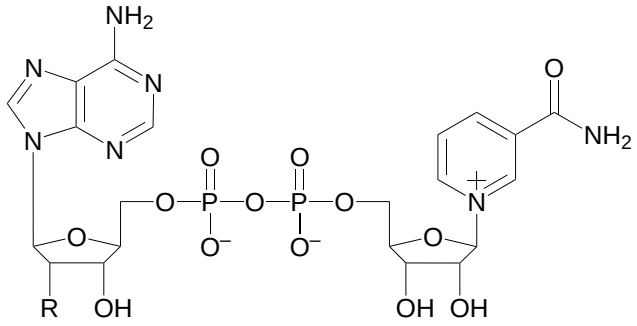
Finally, it is interesting to note that while the major redox change over the diel cycle was in the NADH/NAD ratio, under infection the major redox change was in the NADPH/NADP ratio (Chapter 4). If the NAD(H) pool serves as a reserve for redox energy over the diel cycle of *Prochlorococcus*, it may be that this mechanism breaks down under the shift from constant light to constant dark or under phage infection.

APPENDIX C

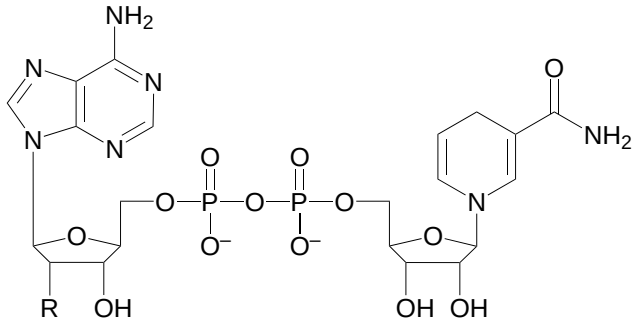
Structures and protocols

Chemical structures

Pyridine nucleotides



$NAD(P)^+$ (oxidized)



NAD(P)H (reduced)

Figure C-1: Structures of pyridine nucleotides. For NAD^+ and NADH, R = –OH. For $NADP^+$ and NADPH, R = $-OPO_3^-$.

Pentose phosphate pathway and Calvin cycle metabolites

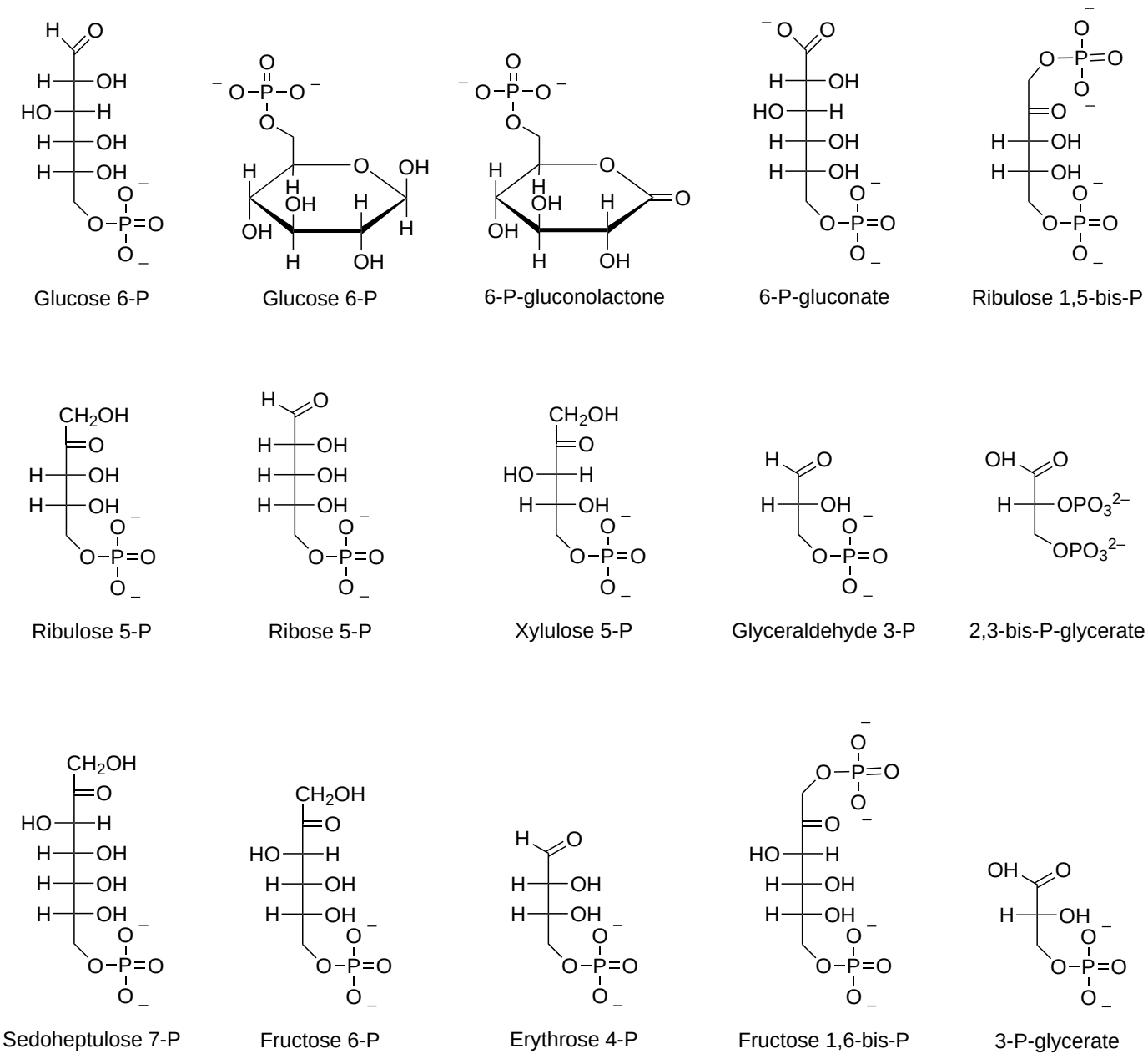


Figure C-2: Structures of pentose phosphate pathway and Calvin cycle metabolites.

Experimental protocols

Measurement of pyridine nucleotides in *Prochlorococcus*

Protocol adapted from Maciejewska and Kacperska (1987), Leonardo et al. (1996), and Tamoi et al. (2005).

1. Sample approximately 100 mL in 40-mL or 250-mL ultracentrifuge bottles (volume sampled depends on cell concentration)
2. Spin 10 min at 20,000×g, 4°C
3. Decant supernatant
4. Resuspend pellets in small volume of leftover supernatant then pool
5. Aliquot equal volumes into 2 Eppendorf tubes labelled "HCl" and "NaOH"
6. Spin 5 min at max, 4°C
7. Aspirate supernatant
8. Resuspend in 200 μL HCl or NaOH (one each for each biological replicate)
9. Boil 5 min
10. Spin 5 min at max, 4°C
11. Transfer supernatant to beadbeater tubes
12. Flash freeze in liquid nitrogen, then store at –80°C

Assay the extracts according to the flowchart on the following page. For data acquired using a Bio-Rad Ultramark Microplate Reader, analyze with plotBioradNadphS.m, ps2pdf.pl, and bioradData.tex, which are available from the author upon request (luket@alum.mit.edu).

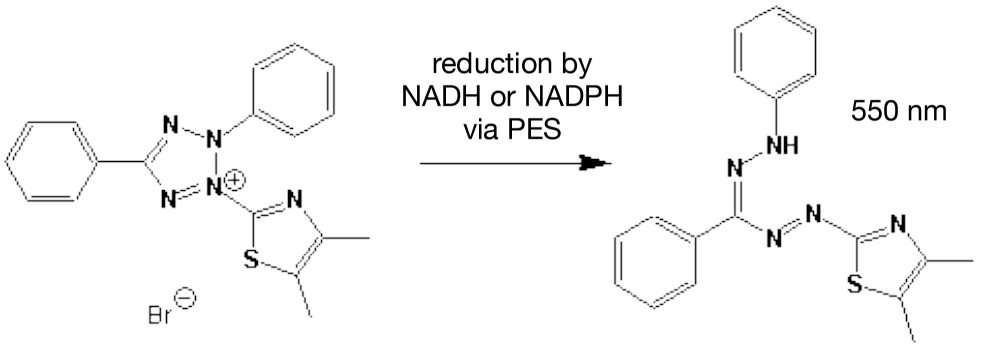
Quantitation of nicotinamide adenine dinucleotides by enzymatic cycling

Harvest cells

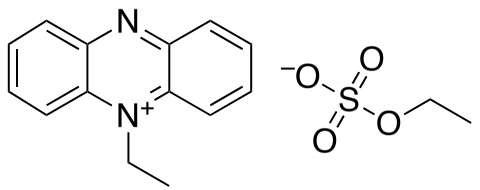
Axenic *Prochlorococcus* or *Synechococcus* culture in mid- to late-log phase
Centrifuge 10 min at 15,000 x *g*, 4°C
Decant and resuspend pellet in ~1/100th initial volume of supernatant
Pool resuspensions and aliquot equally into 2 tubes (1 for HCl, 1 for NaOH)
Centrifuge 5 min at 15,000 x *g*, 4°C
Aspirate supernatant and proceed to extraction

Extraction: Resuspend cell pellets in 200 µL…

NAD⁺/NADP⁺	NADH/NADPH
100 mM HCl, 500 mM NaCl	100 mM NaOH, 500 mM NaCl
Heat 5 min at 95°C, then ice	Heat 5 min at 95°C, then ice
Centrifuge 5 min at 15,000 x *g*, 4°C	Centrifuge 5 min at 15,000 x *g*, 4°C
Transfer supernantat to fresh tube, flash freeze, and store at –80°C	Transfer supernantat to fresh tube, flash freeze, and store at –80°C



MTT_{ox} (tetrazolium, yellow) MTT_{red} (formazan, blue)



PES_{ox} (light-sensitive)

NAD⁺/NADH

Combine 20 µL sample or standard with 180 µL master mix preincubated at 30°C:

100 mM Bicine (pH 8.0), 4 mM EDTA (20 µL 10x) [1400]
10% Ethanol (20 µL 100%) [1400]
1.66 mM Phenazine ethosulfate (20 µL 16.6 mM) [1400]
0.42 mM MTT (20 µL 4.2 mM) [1400]
0.2 U Alcohol dehydrogenase (powder) [14 U]
ddH_2O (100 µL) [7000]

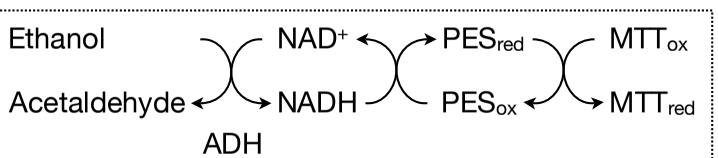
Read absorbance at 550 nm at 30°C for at least 20 min

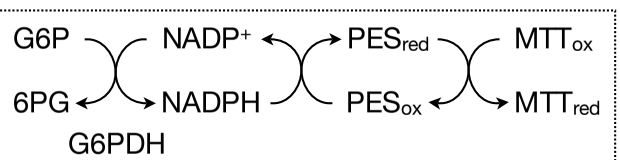
NADP⁺/NADPH

Combine 20 µL sample or standard with 180 µL master mix preincubated at 30°C:

100 mM Bicine (pH 8.0), 4 mM EDTA (20 µL 10x) [1400]
5 mM Glucose-6-P (20 µL 50 mM) [1400]
1.66 mM Phenazine ethosulfate (20 µL 16.6 mM) [1400]
0.42 mM MTT (20 µL 4.2 mM) [1400]
0.2 U Glucose-6-P dehydrogenase (0.2 µL 1 U/µL) [14]
ddH_2O (99.8 µL) [7000]

Read absorbance at 550 nm at 30°C for at least 20 min





Standard curve

20, 50, 100, 200, 500, 1000 nM (final concentrations in assay)
Stocks prepared using A_{340}

Notes

Make master mix for 70 wells (volumes in µL listed in brackets)
Run standards each time
Use normal plates not UV

Quantitative RT-PCR of *Synechococcus* and phage genes during infection

RNA extraction protocol (adapted from Mini RNA Isolation II Kit, Zymo)

1. Resuspend cell pellet in 100 μL 10-mM Tris-HCl pH 8.0
2. Add 5 μL RNase inhibitor (SUPERASE-In, Ambion, 20 U/μL)
3. Add 0.5 μL lysozyme (Ready-Lyse, Epicentre, 30 kU/μL) and mix
4. Incubate 37°C, 30 min
5. Add 0.5 μL lysozyme (Ready-Lyse, Epicentre, 30 kU/μL) and mix
6. Incubate 37°C, 30 min
7. Add 3 volumes (330 μL) ZR RNA Buffer
8. Transfer sample to a Zymo-Spin Column in a supplied Collection Tube
9. Centrifuge at high speed for 1 min and discard flow-through
10. Add 350 μL RNA Wash Buffer to the column, centrifuge at high speed for 1 min, and discard flow-through; repeat
11. Transfer column to an RNase-free microcentrifuge tube
12. Add 50 μL DNase/RNase-free Water directly to the membrane of the column, wait 2 min, and centrifuge at high speed for 1 min
13. Use RNA immediately or freeze at –80°C

DNase treatment for RT-PCR (adapted from Turbo DNA-free Kit, Ambion)

1. Take ∼50 μL freshly extracted RNA
2. Add 6 μL 10× Turbo DNase buffer and mix
3. Add 3 μL Turbo DNase I
4. Quickly vortex and spin down
5. Incubate 37°C, 60 min
6. Vortex inactivation slurry until homogenous
7. Immediately add 10 μL inactivation slurry to each sample, vortexing between each
8. Incubate 2–5 min at room temperature, vortexing every minute or so
9. Centrifuge at high speed for 2 min
10. Transfer supernatant to an RNase-free microcentrifuge tube, taking care not to transfer any inactivation slurry

First-strand cDNA synthesis (adapted from iScript cDNA Synthesis Kit, Bio-Rad)

1. Add the following components to a nuclease-free PCR tube for a 40-μL reaction:

Nuclease-free water	10 μL
5× iScript Reaction Mix	8 μL
RNA template ($\sim$10 ng/μL)	20 μL

2. Heat mixture at 65°C for 5 min and quickly transfer to ice for 1 min
3. Spin down and add enzyme:

iScript Reverse Transcriptase	2 μL

4. Gently mix by hand and spin down
5. Heat using the following protocol:

25°C	5 min
42°C	30 min
85°C	5 min

Quantitative PCR (adapted from QuantiTect SYBR Green PCR Kit, QIAGEN)

1. Add the following components to each well of a qPCR plate:

	96-well	384-well
Nuclease-free water	9 μL	5.4 μL
Forward primer (10 μM)	1.25 μL	0.75 μL
Reverse primer (10 μM)	1.25 μL	0.75 μL
QuantiTect SYBR Green RT-PCR Master Mix	12.5 μL	7.5 μL
cDNA template	1 μL	0.6 μL
	25 μL	15 μL

2. Cover with strip caps or adhesive film and spin down
3. Run using the following temperature cycling protocol (Opticon):

Step 1	95°C	15 min
Step 2	95°C	15 s
Step 3	56°C	30 s
Step 4	72°C	30 s
Step 5	Read fluorescence	
Step 6	Go to Step 2 39 times	
Step 7	72°C	5 min
Step 8	50–90°C	Read fluorescence every 1°C

APPENDIX D

Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts (Sullivan et al., *PLoS Biol*, 2006)

Prevalence and Evolution of Core Photosystem II Genes in Marine Cyanobacterial Viruses and Their Hosts

Matthew B. Sullivan[1☯], Debbie Lindell[1☯], Jessica A. Lee[2], Luke R. Thompson[2], Joseph P. Bielawski[3,4], Sallie W. Chisholm[1,2*]

1 Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, **2** Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, **3** Department of Biology, Dalhousie University, Halifax, Nova Scotia, Canada, **4** Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada

Cyanophages (cyanobacterial viruses) are important agents of horizontal gene transfer among marine cyanobacteria, the numerically dominant photosynthetic organisms in the oceans. Some cyanophage genomes carry and express host-like photosynthesis genes, presumably to augment the host photosynthetic machinery during infection. To study the prevalence and evolutionary dynamics of this phenomenon, 33 cultured cyanophages of known family and host range and viral DNA from field samples were screened for the presence of two core photosystem reaction center genes, *psbA* and *psbD*. Combining this expanded dataset with published data for nine other cyanophages, we found that 88% of the phage genomes contain *psbA,* and 50% contain both *psbA* and *psbD*. The *psbA* gene was found in all myoviruses and *Prochlorococcus* podoviruses, but could not be amplified from *Prochlorococcus* siphoviruses or *Synechococcus* podoviruses. Nearly all of the phages that encoded both *psbA* and *psbD* had broad host ranges. We speculate that the presence or absence of *psbA* in a phage genome may be determined by the length of the latent period of infection. Whether it also carries *psbD* may reflect constraints on coupling of viral- and host-encoded PsbA–PsbD in the photosynthetic reaction center across divergent hosts. Phylogenetic clustering patterns of these genes from cultured phages suggest that whole genes have been transferred from host to phage in a discrete number of events over the course of evolution (four for *psbA,* and two for *psbD*), followed by horizontal and vertical transfer between cyanophages. Clustering patterns of *psbA* and *psbD* from *Synechococcus* cells were inconsistent with other molecular phylogenetic markers, suggesting genetic exchanges involving *Synechococcus* lineages. Signatures of intragenic recombination, detected within the cyanophage gene pool as well as between hosts and phages in both directions, support this hypothesis. The analysis of cyanophage *psbA* and *psbD* genes from field populations revealed significant sequence diversity, much of which is represented in our cultured isolates. Collectively, these findings show that photosynthesis genes are common in cyanophages and that significant genetic exchanges occur from host to phage, phage to host, and within the phage gene pool. This generates genetic diversity among the phage, which serves as a reservoir for their hosts, and in turn influences photosystem evolution.

Citation: Sullivan MB, Lindell D, Lee JA, Thompson LR, Bielawski JP, et al. (2006) Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. PLoS Biol 4(8): e234. DOI: 10.1371/journal.pbio.0040234

Introduction

The marine cyanobacteria *Prochlorococcus* and *Synechococcus* are the smallest and most numerous photosynthetic cells in the oceans [1,2]. The abundances of cyanophages (cyanobacterial viruses) that infect these marine cyanobacteria vary over spatial [3–6] and temporal scales [4,7]—patterns shaped by the dynamics of their host cells [4,8]. Cyanophages are double-stranded DNA viruses belonging to three morphologically defined families: Podoviridae, Myoviridae, and Siphoviridae [3–5,9,10]. Among the cyanophages, podoviruses and siphoviruses tend to be very host-specific, whereas myoviruses generally have a broader host range, even across genera [5], and thus are potential vectors for horizontal gene transfer via transduction.

The movement of genes between organisms is an important mechanism in evolution. As agents of gene transfer, phages play a role in host evolution by supplying the host with new genetic material [11–15] and by displacing "host" genes with viral-encoded homologues [16–18]. Phage evolution is in turn influenced by the acquisition of DNA from their hosts [13,19–22] and by the swapping of genes within the phage gene pool [23,24]. Recent evidence suggests that gene flow within the global phage gene pool extends across ecosystems [25–27].

Cyanophage genomes bearing key photosynthesis genes *psbA* and *psbD* provide a notable example of the co-option of "host" genes for phage purposes [13,22,28–30]. The *psbA* and *psbD* genes encode the two photosystem II core reaction

Academic Editor: Nancy A. Moran, University of Arizona, United States of America

Received February 13, 2006; **Accepted** May 11, 2006; **Published** July 4, 2006

DOI: 10.1371/journal.pbio.0040234

Abbreviations: HL, high-light adapted; LL, low-light adapted

* To whom correspondence should be addressed. E-mail: chisholm@mit.edu

☯ These authors contributed equally to this work.

center proteins, D1 and D2 (denoted here as PsbA and PsbD, respectively), found in all oxygenic photosynthetic organisms. It has recently been shown that the phage-encoded *psbA* gene is expressed during infection [31,32]. Because maximal cyanophage production is dependent on photosynthesis [31,33], and the host PsbA protein turns over rapidly [34] and declines during infection [31], expression of these phage-encoded genes likely enhances photosynthesis during infection, thus increasing cyanophage fitness.

If photosynthesis genes indeed provide a fitness advantage to cyanophages, one might expect them to be widespread among cyanophage genomes. Through whole or partial genome sequencing, *psbA* has been documented in three *Prochlorococcus* cyanophages (one podovirus and two myoviruses) and five *Synechococcus* myoviruses, whereas *psbD* was found in only some of these phages [13,29,35]. Neither of these genes is found in the *Synechococcus* P60 podovirus genome [36]. A survey of *Synechococcus* myovirus isolates revealed that at least 37 of them contained *psbA* [29], and this gene has also been found in cyanophage genome fragments in seawater samples [37]. Thus, the presence of *psbA* is a common, but not universal, feature in the cyanophages examined to date, most of which have been *Synechococcus* cyanophages.

Using limited genomic sequence data from one *Synechococcus* and three *Prochlorococcus* cyanophages, we suggested that both *psbA* and *psbD* were transferred as whole genes from host to phage multiple times, but not from phage to host [13]. Subsequently, Zeidner et al. [37] analyzed *psbA* data predominantly from field sequences and suggested that genetic exchanges of segments of the gene (intragenic recombination) may have occurred among host and phage copies in both directions [37]. However, this novel and controversial hypothesis requires further investigation with sequences of known organismal origin and using methodology capable of identifying the recombination partners and the directionality of such potential exchanges.

To better describe and understand the phenomenon of photosynthesis genes in cyanophage, we looked for the *psbA* and *psbD* genes in 33 cultured cyanophage isolates that infect *Synechococcus* or *Prochlorococcus* (or both) and analyzed the sequences of these genes in the context of known host ranges of the phage. This dataset allowed us to address the following questions: (1) How prevalent are both *psbA* and *psbD* in cyanophages that infect *Synechococcus* and/or *Prochlorococcus*? and (2) To what extent have photosynthesis genes, or segments thereof, been moved between and among hosts and phages?

Results/Discussion

Prevalence of the *psbA* and *psbD* Genes in Cyanophages

The *psbA* gene was amplified from 28 out of the 33 cyanophage isolates examined (Table 1). Combining these findings with published results (Table 1), we find that the *psbA* gene is present in 88% of cyanophage isolates examined, including all myoviruses ($n = 32$) and all five *Prochlorococcus* podoviruses included in this study. However, this gene was not detected in *Prochlorococcus* siphoviruses ($n = 2$) and *Synechococcus* podoviruses ($n = 3$), suggesting that there are some combinations of phage family and host genus that do not lead to incorporation of the *psbA* gene into the phage genome. Six additional phages yielded ambiguous results and were excluded from these analyses (see Materials and Methods for details).

When present, the *psbA* gene is likely to be functional, as there is evidence for the conservation of amino acid sequences through purifying selection [13,37], and the gene is expressed during infection [31,32], implying that this gene confers a fitness advantage to the phages that carry it [13,22,29,31]. Sustained photosynthesis is necessary for maximal phage production [31,33,38], and the long latent period of many freshwater and marine cyanophages (8 h or more; [9,31,33,38]) presumably results in energy- and/or carbon-limitation for phage replication. Thus, cyanophage-encoded *psbA* likely serves to boost the photosynthetic performance of the host during infection, thereby increasing phage production. It is perhaps not coincidental that one of the phages that lacks *psbA*, *Synechococcus* podovirus P60 (Table 1), has a latent period of only 1 h (K. Wang and F. Chen, personal communication), which may be too short for *psbA* expression to be beneficial. Latent period information for marine cyanophages, however, is sparse. It is not known for the *Prochlorococcus* siphoviruses that lack *psbA*, and it has only been shown to be >8 h for a single phage strain from each of the *Synechococcus* myoviruses [39] and *Prochlorococcus* podoviruses [31]. Further, theory [40–43] and experiments [44] suggest that latent period length may be a transient property that rapidly evolves in response to changes in host cell densities. Thus, further exploration of this hypothesis requires analysis of the latent period of many more phage isolates under variable host cell concentrations.

The *psbD* gene was amplified from 15 out of the 33 cyanophage isolates examined (Table 1). Again, combining our data with published findings, we observe that *psbD* is found only in isolates that contain *psbA* and only in myoviruses, but not in all *psbA*-containing myoviruses. Only four of 12 *Prochlorococcus* myoviruses (as defined by original host strain of isolation; Table 1) contained *psbD,* whereas this was the case for 17 of 20 *Synechococcus* myoviruses. Although it is possible that differences in the photosystem II reaction center between *Prochlorococcus* and *Synechococcus* exist (such as differences in the rate of PsbD degradation) and could explain the biased distribution of the *psbD* gene among the myoviruses, there is no evidence that this is the case. The breadth of phage host ranges (as operationally defined in Table 1), however, appears to be a reasonably good predictor of whether a phage will contain *psbD*: 17 of 18 broad-host-range phages encode it, whereas only one out of 21 narrow-host-range phages do so (Table 1). Perhaps broad-host-range phages have co-opted both *psbA* and *psbD* to better ensure the formation of a functional PsbA–PsbD protein complex in the host during infection.

Origins and Evolutionary History of *psbA* and *psbD* in Cyanophages

To investigate the origins of photosynthesis genes in phages and their hosts, we conducted phylogenetic analyses (using measures to minimize systematic errors; see Materials and Methods) of host and phage *psbA* and *psbD* sequences, including new sequence data for nine *Synechococcus* hosts *(psbA),* 19 *Synechococcus* and *Prochlorococcus* hosts *(psbD),* and 33 phages (both *psbA* and *psbD*). Phylogenetic reconstructions of host *psbA* and *psbD* genes in *Prochlorococcus* showed that well-

Table 1. Presence or Absence of *psbA* and *psbD* among *Prochlorococcus* and *Synechococcus* Cyanophages

Family	Phage Name	Cross-Infection[a]	Original Host[a]	Number of Known Hosts[b]	Host Range Breadth[d]	*psbA*	*psbD*	Genome Sequence Confirmation[e]	Reference for *psbA* and *psbD* Sequence
Podoviridae	P-SSP3		*Prochlorococcus* MIT9312	2	Narrow	+	−	−	This study
	P-SSP5		*Prochlorococcus* MIT9515	1	Narrow	+	−	−	This study
	P-SSP6		*Prochlorococcus* MIT9515	1	Narrow	+	−	−	This study
	P-SSP7		*Prochlorococcus* MED4	1	Narrow	+	−	Y	[13]
	P-GSP1		*Prochlorococcus* MED4	1	Narrow	+	−	−	This study
	Syn12		*Synechococcus* WH8017	2	Narrow	−	−	−	This study
	Syn5		*Synechococcus* WH8109	1	Narrow	−	−	Y	This study; P. Weigele, W. Pope, G. Hatfull, R. Hendrix, unpublished data
	P60		*Synechococcus* WH7803	1	Narrow	−	−	Y	[36]
Myoviridae	P-SSM8		*Prochlorococcus* MIT9211	2[c]	Narrow	+	−	−	This study
	P-SSM1		*Prochlorococcus* MIT9303	3	Broad	+	+	−	This study
	P-RSM4		*Prochlorococcus* MIT9303	1[c]	Narrow	+	−	−	This study
	P-SSM2		*Prochlorococcus* NATL1A	3	Narrow	+	−	Y	[13]
	P-RSM5		*Prochlorococcus* NATL1A	1[c]	Narrow	+	−	−	This study
	P-SSM3		*Prochlorococcus* NATL2A	3	Narrow	+	−	−	This study
	P-SSM4		*Prochlorococcus* NATL2A	4	Broad	+, ID to P-RSM2, P-RSM3	+	Y	[13]
	P-SSM9		*Prochlorococcus* NATL2A	2[c]	Narrow	+, ID to P-SSM12	−	−	This study
	P-SSM10		*Prochlorococcus* NATL2A	1[c]	Narrow	+	−	−	This study
	P-SSM12		*Prochlorococcus* NATL2A	2[c]	Narrow	+, ID to P-SSM9	−	−	This study
	P-RSM2	Δ	*Prochlorococcus* NATL2A	6	Broad	+, ID to P-RSM3, P-SSM4	+	−	This study
	P-RSM3	Δ	*Prochlorococcus* NATL2A	6	Broad	+, ID to P-RSM2, P-SSM4	+	−	This study
	S-SM1		*Synechococcus* WH6501	2	Narrow	+	+	−	This study
	S-ShM1		*Synechococcus* WH6501	2	Narrow	+	−	−	This study
	S-SSM1		*Synechococcus* WH6501	2	Narrow	+	−	−	This study
	syn33	Δ	*Synechococcus* WH7803	8	Broad	+	+	−	This study
	S-WHM1	Δ	*Synechococcus* WH7803	5	Broad	+	+	Y	[29]
	S-PM2		*Synechococcus* WH7803	2	Broad	+	+	Y	[29]
	S-RSM2	na	*Synechococcus* WH7803	Unknown	N.D.	+	+	Y	[29]
	S-BM4	na	*Synechococcus* WH7803	Unknown	N.D.	+	+	Y	[29]
	S-RSM88	na	*Synechococcus* WH7803	Unknown	N.D.	+	+	Y	[29]
	syn9	Δ	*Synechococcus* WH8012	13	Broad	+	+	Y	This study; P. Weigele, W. Pope, G. Hatfull, R. Hendrix, unpublished data
	syn10	Δ	*Synechococcus* WH8017	7	Broad	+, ID to syn26	+	−	This study
	syn26	Δ	*Synechococcus* WH8017	9	Broad	+, ID to syn10	+	−	This study
	S-SSM3	Δ	*Synechococcus* WH8018	5[c]	Broad	+, ID to S-SSM5	+	−	This study
	syn30	Δ	*Synechococcus* WH8018	7	Broad	+	+	−	This study
	syn1		*Synechococcus* WH8101	4	Broad	+	+	−	This study
	S-ShM2	Δ	*Synechococcus* WH8102	9	Broad	+	+	−	This study
	S-SSM2	Δ	*Synechococcus* WH8102	9	Broad	+	+	−	This study
	S-SSM5	Δ	*Synechococcus* WH8102	6[c]	Broad	+, ID to S-SSM3	+	−	This study
	S-SSM6	Δ	*Synechococcus* WH8109	7[c]	Broad	+	−	−	This study
	syn19	Δ	*Synechococcus* WH8109	9	Broad	+	+	−	This study
Siphoviridae	P-SS1		*Prochlorococcus* MIT9313	1	Narrow	−	−	−	This study
	P-SS2		*Prochlorococcus* MIT9313	1	Narrow	−	−	−	This study

Presence is indicated by +, and absence by −.

Phages that contained identical sequences to other phages are noted as "ID to X."

[a]Cultured strain used for isolation of phage from natural seawater samples. Phages are defined as either *Prochlorococcus* or *Synechococcus* phages based on original host of isolation, but many of the myoviruses cross-infect both genera. Those phages that cross-infect both genera are marked "Δ", those that do not are left blank, and those that were not available for testing are marked "na".

[b]The number of host strains infected by each phage out of 21 strains tested [5].

[c]Phages whose host ranges are first reported here: P-SSM8, *Prochlorococcus* MIT9211 and MIT9303; P-RSM4, *Prochlorococcus* MIT9313; P-RSM5, *Prochlorococcus* NATL1A; P-SSM9, *Prochlorococcus* NATL1A and NATL2A; P-SSM10, *Prochlorococcus* NATL2A; P-SSM12, *Prochlorococcus* NATL2A and NATL1A; S-SSM3, *Prochlorococcus* NATL1A and NATL2A, *Synechococcus* WH7803, WH8102 and WH8109; S-SSM5, *Prochlorococcus* MIT9303 and MIT9313, *Synechococcus* WH7803, WH8102, WH8103, and WH8109; S-SSM6, *Prochlorococcus* strains NATL1A, MIT9215, and MIT9211, *Synechococcus* strains WH6501, WH8017, WH8018, and WH8109.

[d]Operationally defined as narrow if a phage infects less than four hosts within a single cluster determined from 16S-23S rRNA internal transcribed spacer clustering [48] and broad if it infects more than four hosts within a cluster or at least two hosts that span more than one cluster. Small variations in this definition did not significantly affect the conclusions made.

[e]Indicates whether the PCR results were corroborated by genome sequencing. In all cases where the genome sequence became available, it confirmed the PCR results.

DOI: 10.1371/journal.pbio.0040234.t001

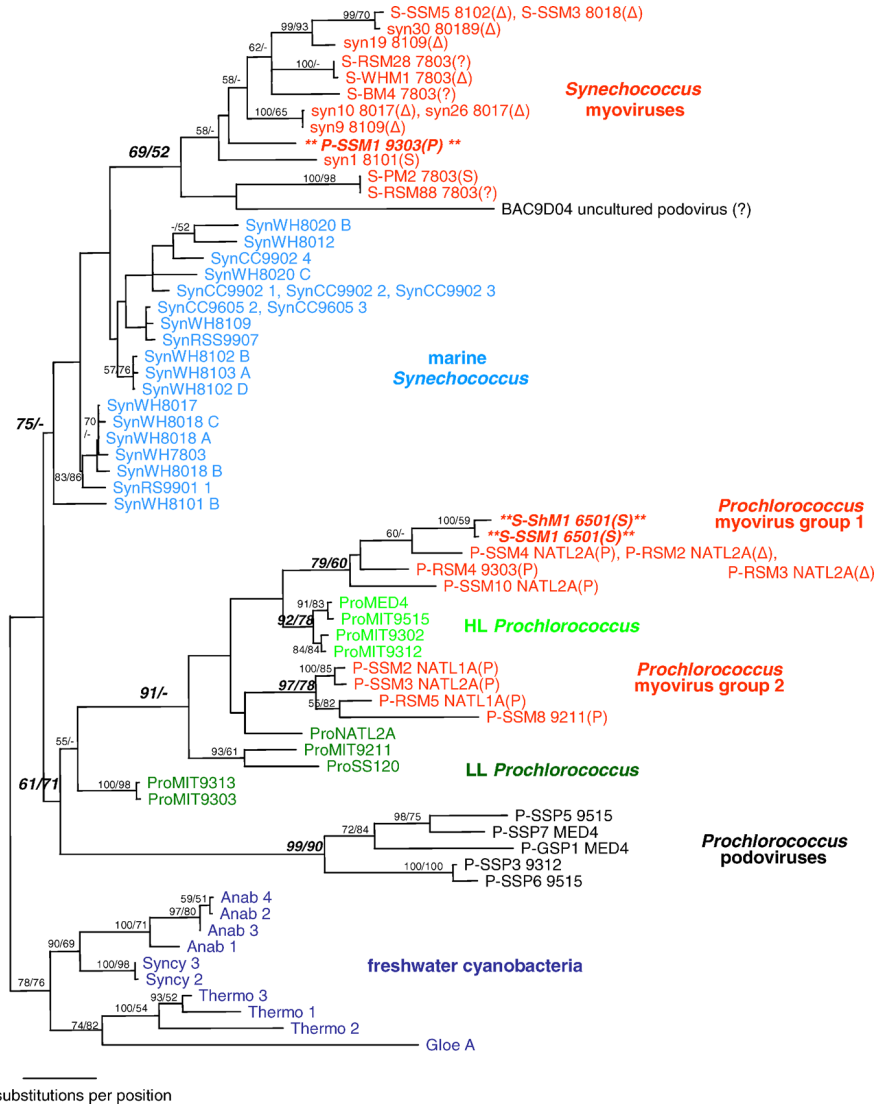


Figure 1. Phylogenetic Tree of *psbA* Gene Sequences from Cultured Cyanobacteria and Cyanophages

Phages are listed by their name, followed by their original host. Phages that are known to infect both *Prochlorococcus* and *Synechococcus* hosts are indicated with a "Δ"; those that infect only one genus are labeled either P (infect only *Prochlorococcus* hosts) or S (infect only *Synechococcus* hosts), while those that are unknown are designated with a "?". Phages shown in italics and bracketed with "**" were isolated on hosts that do not belong to the same cluster and are thus exceptions to the general clustering pattern (see text). Taxa are color coded according to the following biological groupings: myoviruses (red), podoviruses (black), marine *Synechococcus* hosts (light blue), marine *Prochlorococcus* hosts (dark green, LL; light green, HL), freshwater cyanobacteria (dark blue). The tree topology was estimated by LogDet analysis of 1st and 2nd codon positions. Sequences where intragenic recombination was detected using other methods (see Materials and Methods) were not included in these phylogenetic analyses. Branch lengths were estimated by maximum likelihood under a model with nonstationary nucleotide frequencies. Numbers at the nodes represent neighbor-joining bootstrapping and maximum likelihood puzzling support. Anab, *Anabaena;* Gloe, Gleobacter; HL, high-light adapted; LL, low-light adapted; Syncy, *Synechocystis;* Thermo, *Thermosynechococcus*.

DOI: 10.1371/journal.pbio.0040234.g001

supported sequence clusters contain only one organism type (Figures 1 and 2), with sequences from high-light adapted (HL) and low-light adapted (LL) *Prochlorococcus* [45] forming discrete clusters. These well-supported *Prochlorococcus* clusters are similar to those observed using other host genes such as *rRNA*, *rpoC1*, and *ntcA* [46–49], indicating that *psbA* and *psbD* have not been transferred between *Prochlorococcus* lineages. In contrast, the *Synechococcus* clusters for both *psbA* and *psbD* are poorly supported, a finding different to that obtained using other highly conserved genes [46–49] and thus may have resulted from genetic exchange between *Synechococcus* lineages.

The *psbA* sequences from *Synechococcus* myoviruses, *Prochlorococcus* myoviruses, and *Prochlorococcus* podoviruses generally formed discrete clusters consistent with their host ranges (Figure 1), suggesting that the transfer of photosynthesis genes from host to phage has been largely limited by host range (but see exceptions discussed below). Although many of these phages are capable of infecting both host genera (denoted as "Δ" in all figures), we designated each cyanophage isolate as a *Prochlorococcus* or *Synechococcus* cyanophage based upon its original host strain of isolation (as mentioned above and in Table 1). Given this designation scheme, it appears that transfers were predominantly from *Prochlorococcus* to their phages and from *Synechococcus* to their phages. This suggests host-range-limited host-to-phage transfer events,

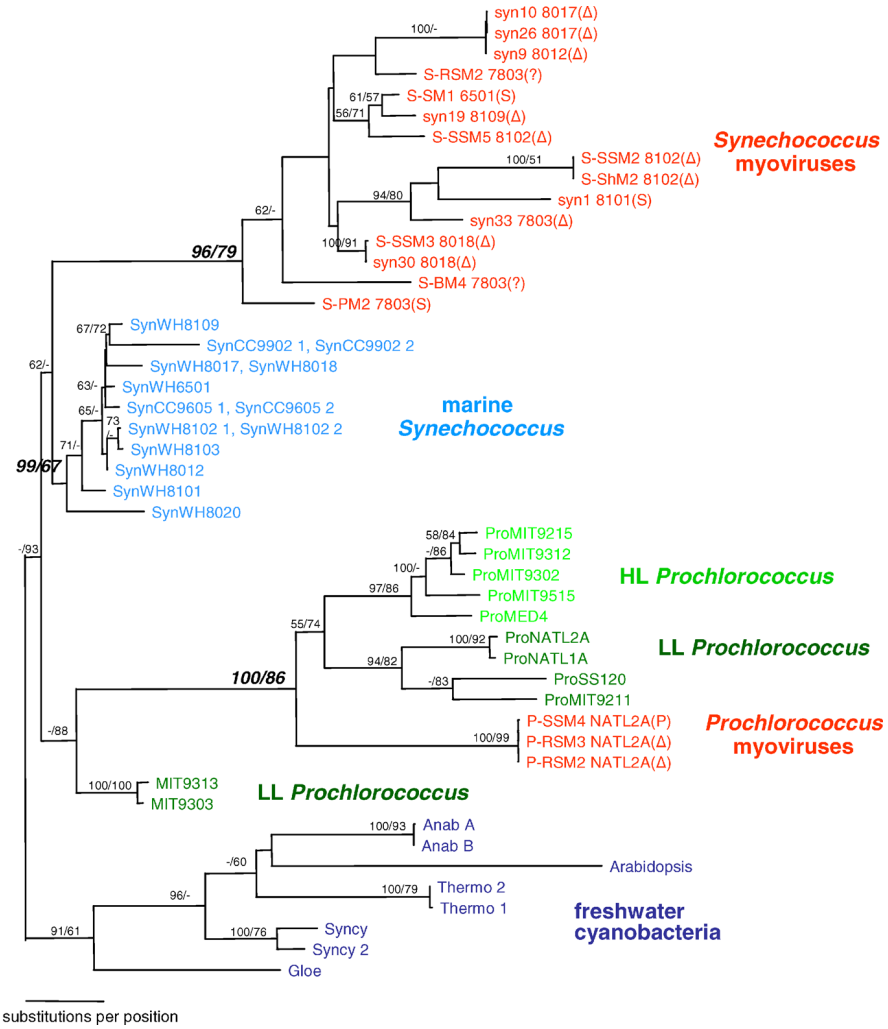


Figure 2. Phylogenetic Tree of *psbD* Gene Sequences from Cultured Cyanobacteria and Cyanophages

Details as in Figure 1. Sequences where intragenic recombination was detected using other methods (e.g., P-SSM1) were not included in these phylogenetic analyses.

DOI: 10.1371/journal.pbio.0040234.g002

with subsequent horizontal and vertical transfers occurring among viral lineages.

Two isoforms of the PsbA protein are often found in cyanobacteria [50]. The PsbA.1 (D1.1) isoform is constitutively expressed, whereas the PsbA.2 (D1.2) isoform is upregulated in response to high light and UV stress [51,52]. Many of the differences between the isoforms are found in ten amino acids between position 121 and 312 [50]. Based on which isoform the majority of these ten amino acids were identical to (including glutamine/glutamate at position 130), we determined that PsbA from both *Prochlorococcus* myoviruses and podoviruses are more similar to PsbA.1, the only isoform found in *Prochlorococcus* hosts so far [53] (unpublished data). Although *Synechococcus* hosts encode both isoforms (unpublished data), *Synechococcus* myoviruses encode the stress-responsive PsbA.2 isoform exclusively (unpublished data), which may be particularly beneficial during the stress of infection. These findings are consistent with the hypothesis of host-range-limited transfers of the *psbA* gene (but see exceptions below).

Host-to-phage transfers appear to have occurred at least four times for *psbA* and twice for *psbD*, as seen from the number of discrete clades containing phage-encoded genes in each case (Figures 1 and 2). The four *psbA* gene acquisitions by phage appear to include two transfer events for the *Prochlorococcus* myoviruses (*Prochlorococcus* myovirus group 1 and 2 in Figure 1) and a single event for *Prochlorococcus* podoviruses all from their *Prochlorococcus* hosts, as well as a single event for *Synechococcus* myoviruses from their hosts (Figure 1). The *psbD* gene appears to have been acquired once by both *Synechococcus* and *Prochlorococcus* myoviruses from their respective hosts (Figure 2). Interestingly, the three *Prochlorococcus* myoviruses that contain *psbD* all encode *Prochlorococcus* myovirus group 1 *psbA* sequences, suggesting that this gene was acquired only once by a subset of these myoviruses. Although the specific source is difficult to determine from phylogeny alone, the placement of the *Prochlorococcus* myovirus sequence clusters suggests that *psbA* was derived from either HL *Prochlorococcus* hosts or LL NATL2A-type hosts, while the *psbD* genes could have been acquired from any of the *Prochlorococcus* hosts other than MIT9313/9303. The placement of the *Prochlorococcus* podovirus (*psbA* only) and *Synechococcus* myovirus sequence clusters at the base of the

host and virus clades provides little further information about the source of these phage genes.

We found three exceptions to the above host-constrained evolutionary scenario—i.e., cases where phage *psbA* and *psbD* genes did not cluster with those of their hosts (Figure 1 and Figures S1 and S2) and did not have PsbA isoforms consistent with that of their hosts (unpublished data). These include two narrow host-range *Synechococcus* myoviruses (S-ShM1, S-SSM1), which encode *psbA* sequences most similar to *Prochlorococcus* myoviruses (Figure 1) even to the extent that they encode the PsbA.1 isoform, as well as a *Prochlorococcus* myovirus (P-SSM1) with a *psbA* sequence that is most similar to those from *Synechococcus* myoviruses (Figure 1) and encodes the PsbA.2 isoform as expected for a *Synechococcus* myovirus. Although the latter can cross-infect across *Prochlorococcus* ecotypes, it has not been shown to infect *Synechococcus* [5]. The P-SSM1 phage also encodes *psbD,* which, like its *psbA* gene, is more similar to *Synechococcus psbD* sequences than those of the *Prochlorococcus* host upon which it was isolated (Figure S2; note that this sequence does not appear in Figure 2 because it was a candidate for intragenic recombination; see Materials and Methods). It is likely that these exceptions to the rather consistent host-phage sequence clustering resulted from horizontal transfer events between a broad-host-range donor phage and a limited-host-range recipient phage during coinfection of a single host, i.e., swapping of genes within the phage gene pool [24]. Whole gene transfers within the phage gene pool are likely to be more common than this, but undetectable when occurring within phages that form a discrete phylogenetic cluster. These observations call for caution when using clustering patterns of *psbA* and *psbD* sequences from uncultured phage (obtained from environmental genome data) to identify potential hosts.

Intragenic Recombination within Core Reaction Center Proteins

The lack of well-supported clade structure in phylogenetic reconstructions for *Synechococcus* host strains when using both *psbA* and *psbD* differs from those constructed using other genes [46–49], which led us to wonder about underlying mechanisms that could be responsible for such a blurred phylogenetic signal. In a recent study, Zeidner et al. [37] showed that *Synechococcus*-phage-like *psbA* sequences from the environment had a patchy %G+C distribution, which they suggest is due to intragenic recombination [37]. Their analyses demonstrated that such recombination had occurred within the inferred-phage clusters and within clusters spanning both phage and host *psbA* sequences. They could not discern, however, whether the signal was caused only by phage-to-phage exchanges, or included phage-to-host exchanges, because the majority of their sequences were of unknown origin (i.e., they were derived from environment clone libraries), and the test employed does not assess the directionality of intragenic recombination events. Our cultured hosts and phages provide an opportunity to assess recombination partners without ambiguity regarding the source of the genes. In addition, the known host ranges of these phages [5] (Table 1), together with the types of recombination tests we have used (see Materials and Methods), allow us to assess the directionality and the pathways through phages and hosts that these recombination events are likely to have taken.

As a first assessment for potential intragenic recombination, we analyzed the %G+C patterns in all of the *psbA* and *psbD* genes (Figures 3 and 4, respectively). *Prochlorococcus* phage genes had similar average %G+C contents to those from their *Prochlorococcus* hosts (39%–46%), whereas those of *Synechococcus* phages had %G+C contents that were lower than those from their *Synechococcus* hosts (46%–51% versus 56%–62%), but not as low as those from *Prochlorococcus* hosts and phages. This intermediate %G+C could be the result of intragenic recombination between variants of the two host lineages. Alternatively, it may reflect the current state of mutational amelioration of the acquired gene from a high %G+C source towards the low genome-wide %G+C of the virus (*Synechococcus* myoviruses S-PM2 and Syn9 both have low genome-wide %G+C; [28]; P. Weigele, W. Pope, G. Hatfull, R. Hendrix, personal communication). If the latter is the case, we might expect such amelioration to be constant across the gene, resulting in an even %G+C distribution pattern.

To help differentiate between these hypotheses, we mapped the %G+C variation across the *psbA* and *psbD* genes using the methodology developed by Zeidner et al. [37]. We detected patchiness of %G+C in *Synechococcus* myovirus *psbA* sequences dispersed along the length of the gene (Figure 3), confirming the findings reported by Zeidner et al. [37]. We also detected %G+C patchiness among *psbA* from *Prochlorococcus* podoviruses, but not from *Prochlorococcus* myoviruses, despite overall similarity of their %G+C content with their *Prochlorococcus* hosts. This suggests that intragenic recombination has occurred among the podoviruses. In addition, patterns of %G+C were not uniform and even markedly clumped across the *psbD* gene from *Synechococcus* myoviruses (Figure 4), with the first segment resembling *Synechococcus* hosts and the last segment resembling *Prochlorococcus* hosts and their phages. Thus, intragenic recombination is likely to be at least partly responsible for the intermediate %G+C content in *Synechococcus* myovirus *psbA* and *psbD* sequences.

Statistical methods for detecting intragenic recombination (see Materials and Methods) revealed strong evidence for its presence in both the *psbA* and *psbD* sequence sets (Tables S1 and S2), but the relative frequency of recombination events was not equal for different groups of hosts and phages. Recombination appears most common among the cyanophages, and more so for *Synechococcus* than *Prochlorococcus* phages. Exchanges were detected between phages that infect both *Synechococcus* and *Prochlorococcus* as well as within myoviruses that infect a single genus *(Synechococcus).* Note that exchanges within a single phylogenetic phage cluster, such as within the *Synechococcus* myoviruses, were undetectable by our previous phylogenetic analyses. Interestingly, our analyses also revealed exchanges between *Prochlorococcus*-specific podoviruses and broad-host-range *Synechococcus* myoviruses, with the *Prochlorococcus* podoviruses serving as the donors (Table S1). Marine cyanobacterial podoviruses contain integrase genes and are thought to have the ability to integrate into the genomes of their hosts as prophages [30] (P. Weigele, W. Pope, G. Hatfull, R. Hendrix, personal communication). If true, genetic exchange could occur between the *Prochlorococcus* prophage and a *Synechococcus* lytic phage—a scenario well accepted in other phage-host systems for genetic exchange [14,15].

Intragenic recombination involving host genes appears less common than phage-to-phage recombination events (Tables S1 and S2). Exchanges between *Synechococcus* and their viruses

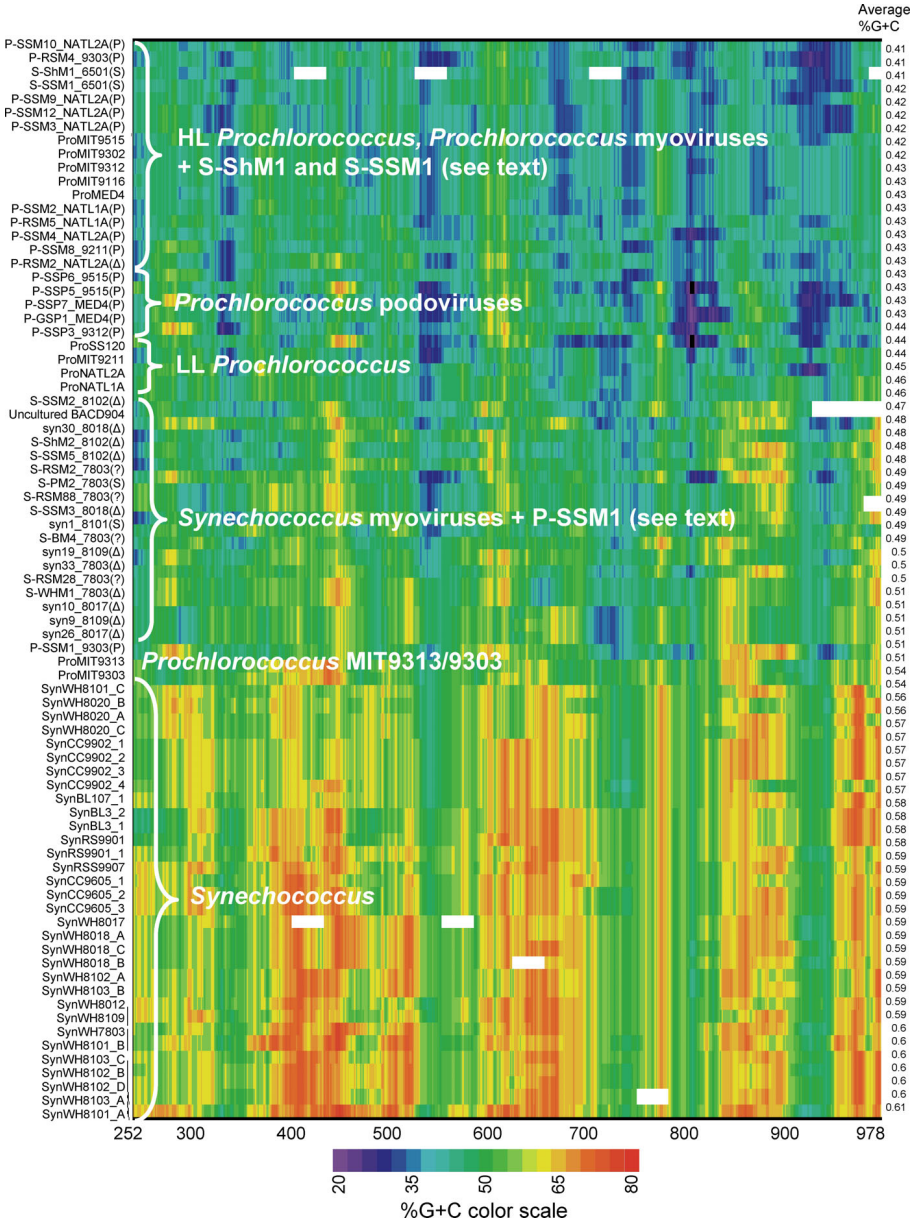


Figure 3. Visualization of %G+C Content across the *psbA* Gene

Colors represent the averaged %G+C in sliding windows along the length of the gene (20%–80%); white regions represent windows that included ambiguous bases in which %G+C could not be calculated for that region. The average %G+C content of the amplified sequence is tabulated on the right side of the figure. Phages are listed by phage name followed by their original host. Phages that are known to infect both *Prochlorococcus* and *Synechococcus* hosts are indicated with a "Δ"; those that infect only one genus or the other have no marker, while those that are unknown are designated with a "?". Host names are prefaced with Syn or Pro for *Synechococcus* and *Prochlorococcus* hosts, respectively. Scale indicates nucleotide positions relative to the *psbA* gene sequence in *Thermosynechococcus*.

DOI: 10.1371/journal.pbio.0040234.g003

are evident, however, and appear to have occurred both from host to phage and phage to host for both *psbA* and *psbD*. Although such events were not detected between *Prochlorococcus* and their phages, there were cases where *Prochlorococcus* myoviruses were the recipients of external DNA from an unknown source (i.e., recombination events possibly involving donors outside of our dataset). Thus, phages may be contributing to the intragenic recombination of portions of these genes in *Synechococcus,* perhaps explaining the lack of phylogenetic structure observed in *psbA* and *psbD* trees for *Synechococcus* clusters (but not for *Prochlorococcus* clusters) relative to those obtained when using other phylogenetic markers [46–49]. Presumably, phage-host intragenic exchanges occur via homologous recombination during infection. Clearly, the transfer of DNA will be retained in host lineages only if infection fails to lyse the host (e.g., abortive infection [54]).

Finally, intragenic exchanges among hosts were also occasionally detected, particularly among *Synechococcus* (Tables S1 and S2). This may also play a role in the lack of clade structure among *Synechococcus* strains in the *psbA* and *psbD* trees. Although two possible intragenic recombination events between *Synechococcus* and *Prochlorococcus* were identified, they were resolved as small regions (15–16 bases) and may be false

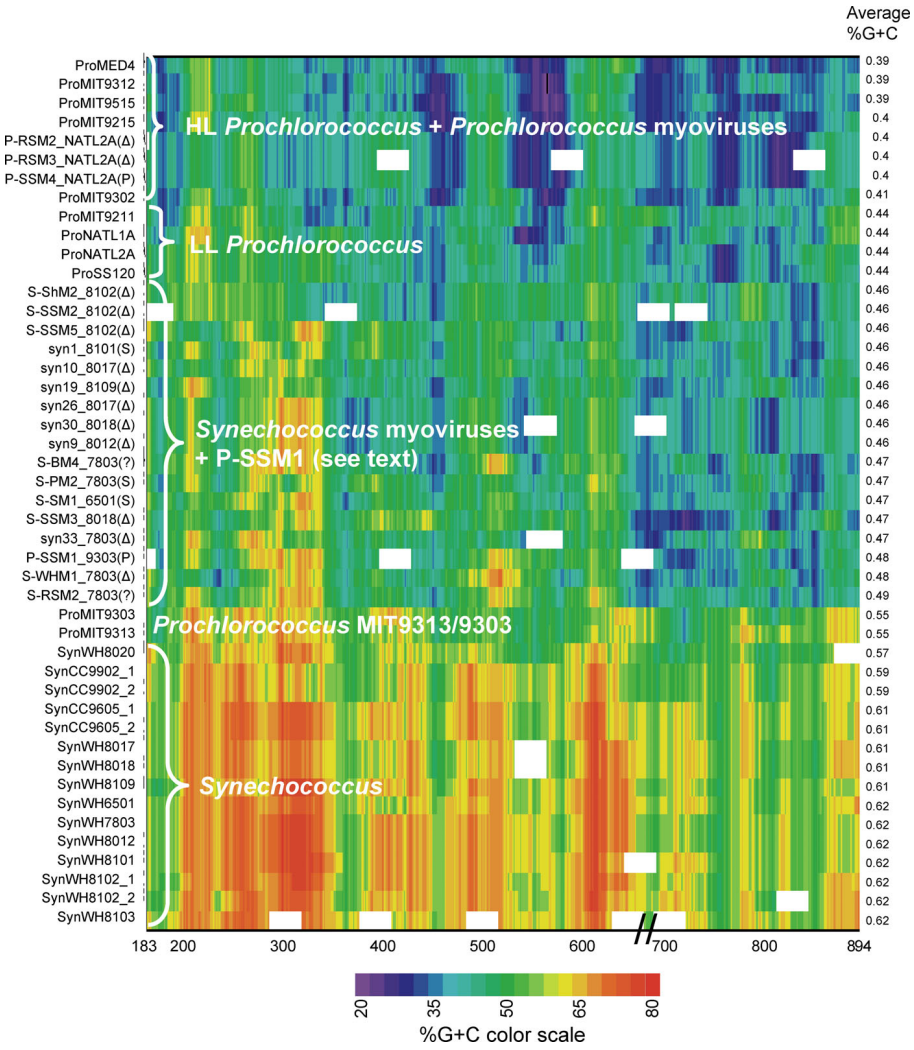


Figure 4. Visualization of %G+C Content across the *psbD* Gene

Details as in Figure 3. Note that the 21-nucleotide indel in *Prochlorococcus* hosts and their phages [13] (unpublished data) was excluded from the analysis at the position indicated by the "//" symbol to maximize the data that could be displayed using the sliding window approach.

DOI: 10.1371/journal.pbio.0040234.g004

positives. Host-to-host transfers may have occurred through the uptake of DNA directly from the environment (e.g., via transformation) or through viral intermediates [37]. Such host-to-host intragenic exchanges via viral intermediates presumably occur through generalized transduction [55].

In summary, our findings suggest that the shuffling of segments of *psbA* and *psbD* within the cyanophage gene pool has generated significant photosynthesis gene diversity and serves as an extended reservoir of genetic diversity for their hosts, influencing photosystem evolution.

psbA and *psbD* Gene Diversity in Cultured Isolates Captures Most of the Field Diversity

We next sought to determine how well *psbA* and *psbD* sequence diversity observed in culture collections represents that observed in wild phage populations, and whether additional whole-gene host-to-phage transfer events could be identified from these wild sequences from the phage gene pool. Zeidner et al. [37] had previously examined field diversity of the *psbA* gene sequence from environmental samples where *Synechococcus* strains were the dominant phototroph [37]. Thus, we sought to examine genetic diversity of this gene, as well as that of *psbD,* from an environment where *Prochlorococcus* cells commonly outnumber *Synechococcus* cells by orders of magnitude [56]. To this end, we amplified, cloned, and sequenced *psbA* and *psbD* gene sequences obtained from the viral-sized fraction (0.02–0.2 μm) of two seawater samples within (25 m) and below (75 m) the mixed layer in the Pacific Ocean off the coast of Hawaii (Figures 5 and 6, respectively). The *psbA* and *psbD* sequences from these viral-fraction samples clustered with cultured *Prochlorococcus* cyanophage isolates (with varying levels of support; Figures 5 and 6), but not with *Synechococcus* cyanophages. There was not a notable difference in the phylogenetic placement of the *psbA* or *psbD* clones obtained from within or below the mixed layer. Although this suggests a lack of vertical structure in diversity among the sequence types, we did not sequence these samples to saturation; thus, such conclusions are preliminary.

More than half of the wild *psbA* sequences (42 of 81) form a large cluster with cultured *Prochlorococcus* podoviruses (Figure 5). Within this group, all but one cluster of wild sequences

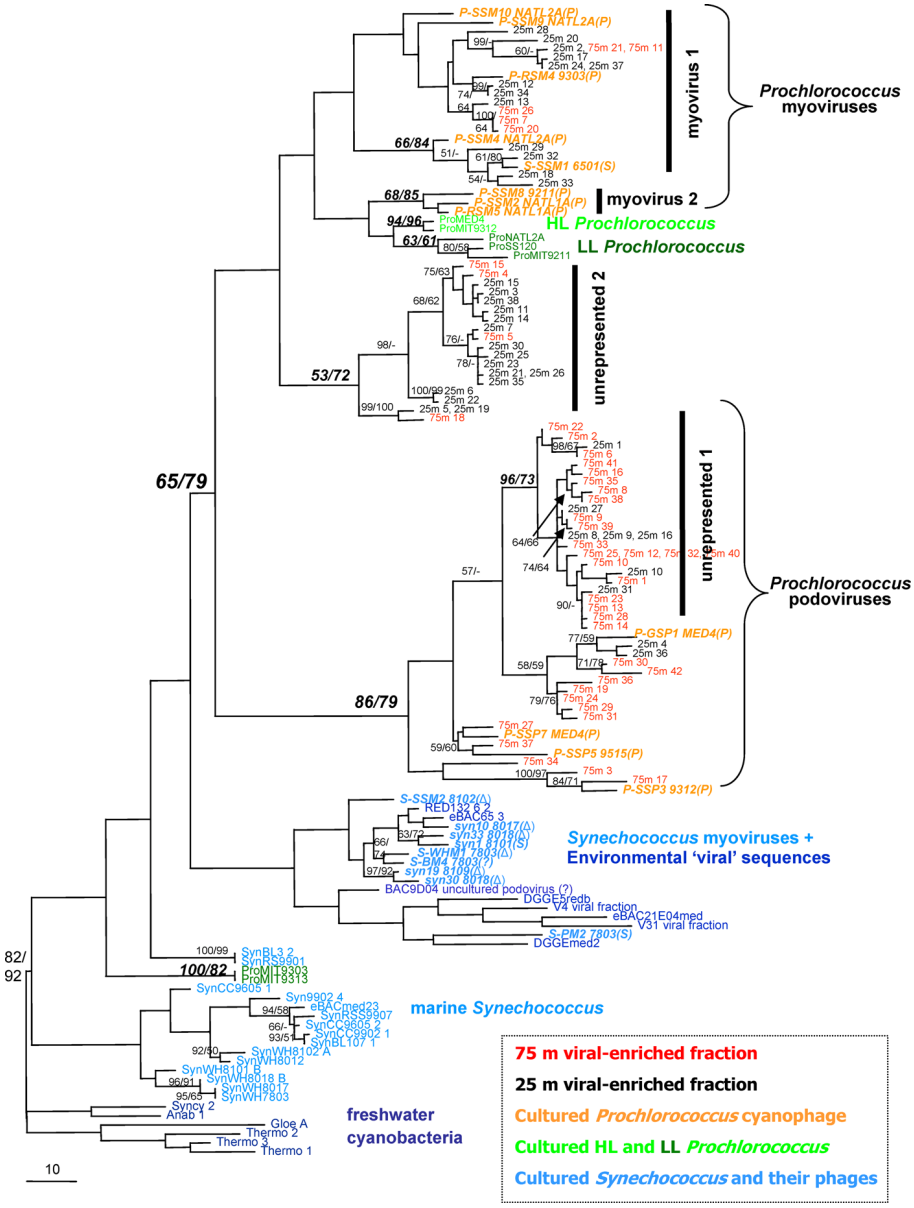


Figure 5. Phylogenetic Tree of *psbA* Gene Sequences from Representative Cultured Cyanobacterial and Cyanophage Isolates and Cloned Environmental Sequences from the Hawaii Ocean Time Series Site in the Pacific Ocean

Phylogenetic tree of *psbA* gene sequences and cloned environmental sequences were collected from above (25 m, black) and below (75 m, red) the surface mixed layer at the Hawaii Ocean Time Series site in the Pacific Ocean, a region where *Prochlorococcus* are the dominant phototrophs. Details for naming conventions are as in Figure 1. *Synechococcus* environmental "viral" sequences from [37]. The tree topology was estimated by LogDet analysis of 1st and 2nd codon positions, with branch lengths estimated using stationary nucleotide frequencies.

DOI: 10.1371/journal.pbio.0040234.g005

contain cultured podovirus sequences (Figure 5). The extensive microdiversity in this cluster (labeled "unrepresented 1") was probably derived from within the podovirus gene pool, as evidenced by the presence of podovirus phage isolates in the more basal branches of the cluster. Other *psbA* sequences from the field samples form subclusters that contain cultured *Prochlorococcus* myoviruses and form a large group that also contains *Prochlorococcus* hosts (Figure 5). One cluster ("unrepresented 2" in Figure 5) within this group also lacks sequences from cultured hosts or phages. The basal position of this cluster suggests that these sequences may belong to phages that infect as-yet uncultured *Prochlorococcus* hosts [57] and may represent an additional host-to-phage transfer event. Thus, our work here, together with that of Zeidner et al. [37], suggests that cyanophage culture collections represent much of the naturally occurring *Prochlorococcus* and *Synechococcus* cyanophage *psbA* gene sequence diversity [37].

All *psbD* sequences from wild phages fall into a single well-supported cluster that includes a representative cultured *Prochlorococcus* cyanophage P-SSM4 (Figure 6). This cluster reveals significant microdiversity within the *psbD Prochlorococcus* phage gene pool in the viral-fraction from this Pacific Ocean site and suggests that phages that encode *Prochlorococcus*-phage-like *psbD* genes are perhaps not rare in this environment. The four *Prochlorococcus* cyanophages that

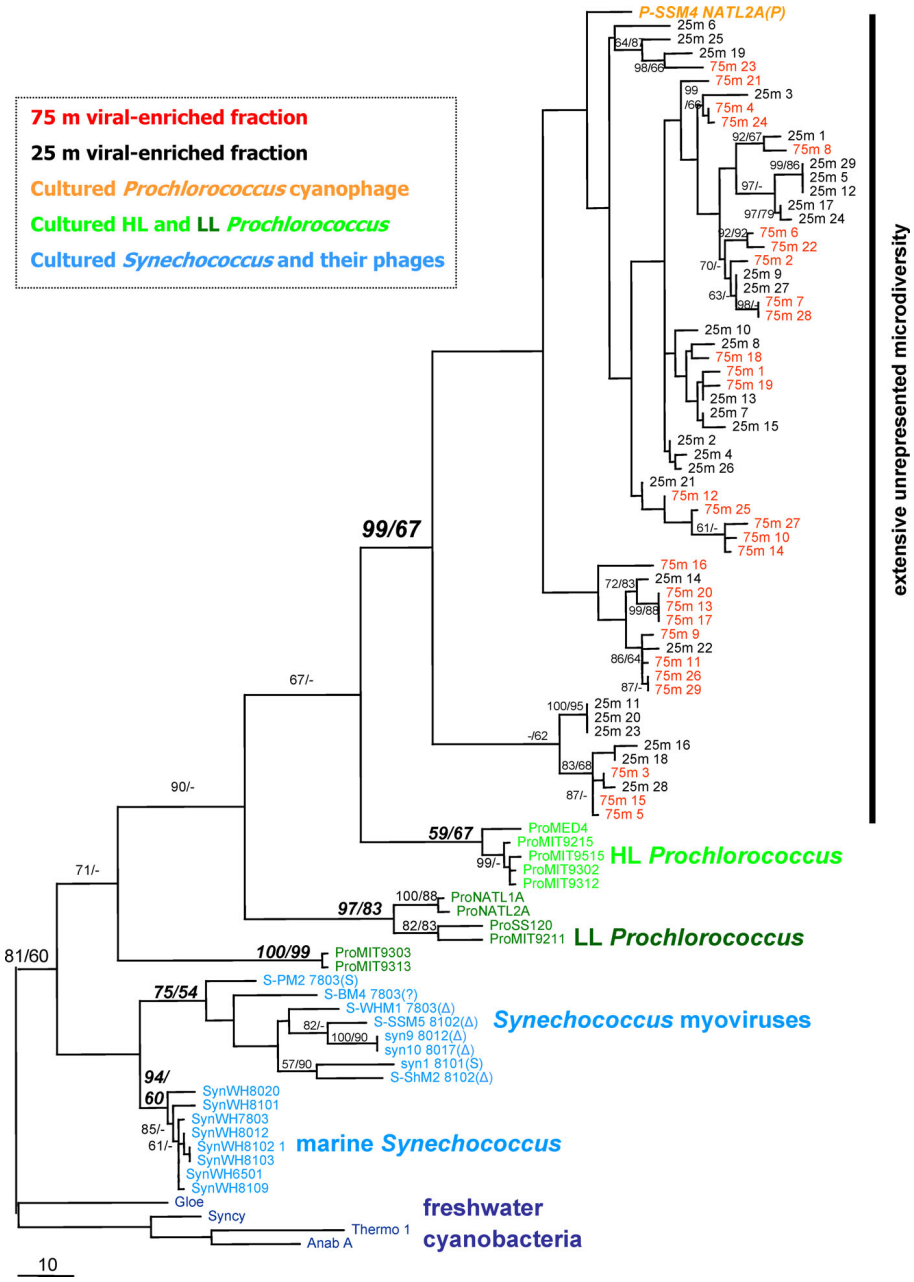


Figure 6. Phylogenetic Tree of *psbD* Gene Sequences from Cultured Cyanobacterial and Cyanophage Representatives and Cloned Environmental Sequences from the Pacific Ocean

Details for naming conventions are as in Figure 1, and phylogenetic analyses are as in Figure 5.

DOI: 10.1371/journal.pbio.0040234.g006

contain the *psbD* gene in our culture collection originated from either the Sargasso Sea or the Red Sea; thus, it is perhaps not surprising that the viral-fraction microdiversity from the Pacific Ocean is largely unrepresented in this collection.

Conclusions

The phage genomic repertoire evolves through the exchange of genetic material from other phages [24] and by co-opting metabolic genes from their hosts [13,20,22]. The prevalence of photosynthesis genes in cyanophages strongly suggests that the capture of these genes provides a significant fitness advantage among certain cyanophage types. Previously, we have shown that the horizontal transfer of *hli* genes from cyanophages to their hosts has likely played a role in driving host niche differentiation [13]. More recently, cyanophages were hypothesized to be involved in partial gene exchanges even for the core photosystem gene *psbA* of their hosts [37]. Here, we show that genetic exchanges involving cyanophages may have influenced the make-up of both of the core photosystem II genes *(psbA* and *psbD)* in *Synechococcus,* whereas this was less apparent for *Prochlorococcus*. Therefore, mounting evidence indicates that host-like genes acquired by phages undergo a period of diversification in

phage genomes and serve as a genetic reservoir for their hosts. Thus, a complex picture of overlapping phage and host gene pools emerges, where genetic exchange across these pools leads to evolutionary change for host and phage. Fully understanding the mechanisms of microbial and phage coevolution clearly requires an improvement in our ability to quantify horizontal gene transfer at the whole and partial gene level and in our ability to accurately estimate the relative fluxes into and out of these pools.

Materials and Methods

DNA isolation from cultured hosts and phages and environmental samples. Eleven strains of *Prochlorococcus,* ten strains of *Synechococcus,* and 38 phages of *Prochlorococcus* and *Synechococcus* (seven podoviruses, 29 myoviruses, and two siphoviruses) were screened for *psbA* and *psbD* sequences for this study. We report here on new *psbA* sequences from nine *Synechococcus* hosts and new *psbD* sequences from 19 *Prochlorococcus* and *Synechococcus* hosts (including two from unpublished *Synechococcus* genomes for strains CC9605 and CC9902; available from http://genome.jgi-psf.org/mic_home.html. The 38 phages screened included seven phage templates for which genome sequences are now available (P-SSM2, P-SSM4, P-SSP7, S-PM2, S-WHM1, Syn5, Syn9), enabling us to validate our PCR amplification findings. Host genomic DNA was extracted using a DNeasy Tissue Kit (Qiagen, Valencia, California, United States). Filtered (0.2 μm, Acrodisc supor membrane syringe filter) phage lysates in Pro99 medium were used as DNA templates for subsequent PCR amplification experiments.

Environmental samples were collected from the Hawaii Ocean Time Series (HOT) on 15 October 2003 at 45°N 158°W from depths of 25 m and 75 m. These samples were filtered through a 0.2-μm filter (Osmonics, Minnetonka, Minnesota, United States, Poretics polycarbonate 25-mm filter) to remove cellular material and substantially enrich for environmental phages. A 100-ml volume of 0.2-μm filtrate was then filtered onto a 0.02-μm filter (Whatman Anotop 25) to collect phage particles and resuspended in 7 ml of a modified SM storage buffer (600 mM NaCl, 8 mM $MgSO_4$-$7H_2O$, 50mM Tris [pH 7.5], 0.04% gelatin).

Overview of *psbA* and *psbD* screening strategy. PCR screening for *psbA* and *psbD* across a diverse set of samples presented several challenges. These included variable amplification efficiencies, uncertainty about whether amplicons derived from phage or host, and multiple gene copies in hosts. The amplification strategy was as follows: for each virus and host strain, four PCR reactions were carried out, pooled, and analyzed by gel electrophoresis; if the amplification product was not visible, it was diluted 10-fold and used as template for nested or semi-nested PCR and the resulting products analyzed; if still no product was visible, multiple phage stocks were rescreened. Multiple copies of *psbA* in *Synechococcus* strains were identified by sequencing many clones and were distinguished from sequencing errors as described below. We did not screen for multiple copies of *psbA* from *Prochlorococcus* or multiple copies of *psbD* from either *Synechococcus* or *Prochlorococcus,* as when present, they are generally indistinguishable from each other [58–60].

Amplification of *psbA* and *psbD*. PCR reactions were performed with *Taq* DNA polymerase and deoxyribonucleotide triphosphates from New England Biolabs (Beverly, Massachusetts, United States) or Invitrogen (Carlsbad, California, United States) and carried out with a PTC-100 or PTC-200 DNA Engine (MJ Research, Waltham, Massachusetts, United States) or a Robocycler Gradient 96 (Stratagene, La Jolla, California, United States). Template amounts were 10 ng of genomic DNA for *Prochlorococcus* and *Synechococcus,* 1 μl of lysate for cyanophages, and 2 μl of filtrate for environmental samples. PCR primers and amplification reaction conditions are shown in Tables S3 and S4.

The *psbA* gene from all sources was amplified using primer pair *psbA*-F/R [61] and PCR protocol A (Tables S3 and S4). Four reactions were conducted with each template, and the products were pooled and analyzed by agarose gel electrophoresis. Primer *psbA*-R falls on the intron region in S-PM2 [29]. Therefore, for efficient amplification of phage *psbA* genes that may contain introns, and for increased sensitivity, we used the Pro-*psbA*-F/R primer set and protocol B in nested PCR reactions when no PCR product was visible from cyanophage lysates and environmental filtrates. To reduce the incidence of heteroduplex formation, amplification products from environmental samples were subjected to reconditioning PCR [62]: initial PCR products were diluted 1:10, then amplified using protocol A but for only three cycles.

The *psbD* gene from *Prochlorococcus, Synechococcus,* and cyanophages was amplified using primer pair *psbD*-54F/*psbD*-308R and protocol D. However, when product yield was low or absent, semi-nested PCR was carried out as follows. Amplification was first conducted using primer pair *psbD*-26F/*psbD*-308R and protocol C. Four reactions were conducted with each template, the products were pooled, diluted 1:10, and used as templates for a second round of amplification using primer pair *psbD*-54F/*psbD*-308R and protocol D. *psbD* from environmental samples was amplified using primer pair *psbD*-26F/*psbD*-308R and protocol C and subjected to reconditioning PCR as for *psbA* (see above).

In preparation for sequencing, PCR products were either purified directly using the QIAquick PCR Purification Kit (Qiagen) or separated on an agarose gel and then purified using the QIAquick Gel Extraction Kit (Qiagen).

To confirm that the absence of *psbA* or *psbD* PCR products from phage was not simply due to a lack of amplifiable phage DNA, we screened phage lysates for known phage genes: *g20* (for myoviruses) and *DNApol* (for podoviruses). *g20* was amplified using primer pair g20-F/R and protocol E, and *DNApol* using primer pair DNApol-F/R and protocol F, both with 1 μl of lysate. In all cases, a product was obtained, suggesting the phage template DNA was present and amplifiable by PCR (unpublished data).

Six phage lysates yielded PCR products with sequences identical to those of a known host. These six phage lysates include five cyanophages previously described (P-RSP1, P-SSP1, P-SSP2, P-ShM1, P-ShM2; [5]), as well as one cyanophage not previously reported in the literature (P-SSP9; M.B.S. and S.W.C., unpublished data). In these cases we could not eliminate the possibility that the amplicon resulted from host DNA, the amplification of which may be more likely to occur when there is no phage template for this gene. Thus, we excluded these phages from further analyses. In contrast, phages with amplicon sequences identical to those of other phages (indicated as "ID to X" in Table 1) were passed through multiple lysates, and a "fingerprint" phage gene *(g20)* was used to confirm that there was a single phage in the lysate. The *psbA* sequence was then reassayed, increasing our confidence in these results. Even with this precaution, we cannot rule out the possibility of PCR contamination for those few cases where identical sequences were amplified from different phage lysates.

Cloning and sequencing of PCR products. The *psbA* gene is often found in multiple distinct copies in marine *Synechococcus* [59], whereas in *Prochlorococcus* the *psbA* gene is either single copy per genome or encodes multiple copies that are nearly identical to each other [60,63,64]. Among cyanophages, the *psbA* gene has only been found in a single copy per genome [28,30]. To allow for the identification of multiple *psbA* gene copies in *Synechococcus* strains, PCR products from *Synechococcus* templates were cloned prior to sequencing. Cloning was performed using the TOPO TA Cloning Kit for Sequencing (Invitrogen) with the pCR4-TOPO vector. Ligation products were transformed into TOP10 competent cells. Plasmid purification and sequencing were conducted by Genaissance Pharmaceuticals (New Haven, Connecticut, United States). Inserts were sequenced from both forward and reverse directions, using the M13F and M13R primer binding sites in the pCR4-TOPO vector.

Approximately ten *psbA* clones were sequenced for each *Synechococcus* strain. The published genome of *Synechococcus* WH8102 provides an example of natural *psbA* diversity in a given strain, as it contains four copies of *psbA:* two copies that are 99.8% identical and a third and fourth copy that are 99.4% and 88% identical, respectively, to the above two *psbA* copies [59]. Considering a *Taq* polymerase error rate of 3×10^{-5} per nucleotide per duplication [65], at most one error could be expected in each *psbA* gene sequenced. Thus, sequences were considered identical, and removed from the analysis pool, if they were more than 99.8% identical, to avoid data issues stemming from possible PCR error (sequencing error should be nonexistent because consensus sequences were obtained from forward and reverse sequencing of the clones). Sequence identity levels for nonidentical clones from the remaining dataset ranged from about 60% to 99.0%.

PCR products from genes presumed not to have multiple distinct copies per genome (*psbA* from *Prochlorococcus* and cyanophage; *psbD* from all organisms) were generally sequenced directly (Harvard Medical School Biopolymers Facility [Boston, Massachusetts, United States], Davis Sequencing [Davis, California, United States], or Genaissance Pharmaceuticals). The absence of multiple significant-height peaks at single nucleotide positions in chromatograms from this direct sequencing (unpublished data) confirmed that single products were amplified during PCR. Each strain was sequenced in

both forward and reverse directions, using the same primers used for PCR amplification.

Sequence analyses. Previous analyses have raised important concerns about using *psbA* gene sequence datasets that may suffer from large %G+C variability and conflicting phylogenetic signals in phylogenetic reconstructions [37]. To minimize such errors, we followed these steps.

We first performed phylogenetic analyses using sequences from all taxa (80 for *psbA* and 50 for *psbD*) and all codon positions (Figures S1 and S2). Phylogenetic trees were constructed by using distance and maximum likelihood. Neighbor-joining [66] was used to reconstruct a distance tree under the HKY85 model [67]. Maximum likelihood analysis was performed under HKY85 combined with a gamma model for among sites rate variation, assuming eight rate categories with model parameters estimated from the data [68]. Maximum likelihood trees were obtained by quartet puzzling, as implemented in the program TREE-PUZZLE 5.0 [69]. Bootstrap resampling (1,000 pseudoreplicates) was used to measure the relative support for internal branches of the neighbor-joining trees. For quartet puzzling, support was estimated from 25,000 (*psbD* trees) or 50,000 (*psbA* trees) pseudoreplicates.

These analyses resulted in trees with high bootstrap support at many critical nodes (Figures S1 and S2). However, fitting a single tree to large datasets containing conflicting phylogenetic signals can lead to reconstruction artifacts (i.e., systematic errors) that result in high bootstrap support [70,71]. We found, using neighbor-nets [72] constructed by using the SplitsTree2 program [73], within-gene conflicting phylogenetic signals in both the *psbA* and *psbD* datasets as indicated by the box-like structures in neighbor-nets graphs (Figures S3 and S4). Specifically, networks for both genes revealed substantial conflict involving splits between *Synechococcus* strains, their myoviruses, and a complex of sequences comprised of *Prochlorococcus* and their viruses.

We further investigated whether these large datasets could suffer from systematic errors related to: (i) substitution rate variation among lineages [74], (ii) heterogeneous compositional bias among lineages (e.g., %G+C; [75]), and (iii) within-gene heterogeneity in phylogenetic signals [76]. We found significant substitution rate variation among lineages (Table S5) using likelihood ratio tests. In addition, nucleotide frequencies were nonstationary across these data, with significant differences in equilibrium frequencies for clades defined according to organism types (Table S6; [77]). Not surprisingly, the largest divergence in %G+C across taxa was at the 3rd codon positions of both *psbA* and *psbD*.

Zeidner et al. [37] hypothesized intragenic recombination in *psbA* [37]. We attempted to identify this qualitatively through graphical analysis of %G+C and quantitatively using four different tests for intragenic recombination. The %G+C distribution was examined within overlapping sequence windows (a sliding window of 30 nucleotides with a five-nucleotide step) using the GCViz script [37] (available upon request from Dr. Shmoish of Technion–IIT; mshmoish@cs.technion.ac.il) written in the R-language (http://www.r-project.org). Three of the four different tests for within-gene recombination are based on the distribution of substitutions (GeneConv: [78]; MaxChi: [79]; Chimaera: [80]), while the fourth used a phylogenetic approach ("RDP," as implemented in [81]). We considered only those recombination events that satisfied all of the following criteria: (i) results were significant after application of Bonferroni correction for multiple tests, (ii) regions were detected by two or more different methods, and (iii) consensus breakpoints could be estimated for a given region identified using different methods. Once a putative recombination event was detected, we inferred the best candidate donor sequence (that most similar to the recombinant segment) using RDP [81].

In summary, to minimize systematic errors in the ultimate phylogenetic analyses, we first processed the dataset as follows: (i) excluded those sequences having a strong signal for intragenic recombination, (ii) excluded 3rd codon positions, which display the largest differences in %G+C and substitution rates among lineages, and (iii) employed LogDet distances [75] to accommodate compositional heterogeneity (variable %G+C) in the remaining data. These measures proved to be important. The uncorrected dataset grouped lineages according to evolutionary rates and %G+C bias (Figures S1 and S2), whereas the ultimate analysis did not (see Figures 1 and 2). Statistical analysis of the processed dataset under nonhomogenous evolutionary models [77] revealed that the ultimate phylogenetic hypotheses (see Figures 1 and 2) provided a significantly better fit to the data (Table S7). Prior to processing the data, the alternative phylogenies were indistinguishable (Table S7).

Supporting Information

Figure S1. Phylogenetic Analyses Including All *psbA* Gene Sequences from Cultured Cyanobacteria and Cyanophages

Phages are listed by phage name, followed by their original host. Host range information is designated in parentheses. Phages known to infect both *Prochlorococcus* and *Synechococcus* hosts are indicated with a "Δ"; phages that infect only *Prochlorococcus* or *Synechcococcus* are designated by a P or S, respectively; and those host ranges that are unknown have a "?". Phages shown in italics and bracketed with "**" were isolated on hosts that do not belong to the same cluster and are thus exceptions to the general clustering pattern (see text). Taxa are color coded according to the following biological groupings: myoviruses (red), podoviruses (black), marine *Synechococcus* hosts (light blue), marine *Prochlorococcus* hosts (dark green, HL; light green, LL), freshwater cyanobacteria (dark blue). Neighbor-joining tree was inferred under HKY85 mode and using sequences from all taxa and all codon positions. Nucleotide frequencies were assumed to be homogenous across lineages. Numbers at the nodes represent neighbor-joining bootstrapping and maximum likelihood puzzling support. Anab, *Anabaena;* Gloe, Gleobacter; HL, high-light adapted; LL, low-light adapted; Syncy, *Synechocystis;* Thermo, *Thermosynechococcus*.

Found at DOI: 10.1371/journal.pbio.0040234.sg001 (79 KB PPT).

Figure S2. Phylogenetic Analyses Including All *psbD* Gene Sequences from Cultured Cyanobacteria and Cyanophages

Details are as in Figure S1.

Found at DOI: 10.1371/journal.pbio.0040234.sg002 (59 KB PPT).

Figure S3. Neighbor-Nets Analysis of 80 *psbA* Gene Sequences (including All Cyanophage and Marine Cyanobacterial Sequences Available)

The analysis was conducted under the HKY85 model of substitution using all codon positions. Taxa color coding and abbreviations are as in Figure S1. The box-like appearance in the basal branches of this phylogeny suggests regions of conflicting phylogenetic signals (see Materials and Methods).

Found at DOI: 10.1371/journal.pbio.0040234.sg003 (272 KB PDF).

Figure S4. Neighbor-Nets Analysis of 50 *psbD* Gene Sequences (including All Cyanophage and Marine Cyanobacterial Sequences Available)

Taxa color coding and abbreviations are as in Figure S1. Details of the analysis are as in Figure S3.

Found at DOI: 10.1371/journal.pbio.0040234.sg004 (249 KB PDF).

Table S1. Consensus Results from Four Tests for Intragenic Recombination within Gene Sequences in Our *psbA* Dataset

The four tests included (1) RDP, (2) GeneConv, (3) MaxChi, and (4) Chimaera (as described in Materials and Methods), and recombination was considered "detected" only when the following criteria were satisfied: (i) similar regions were detected by two or more methods, (ii) all such regions were significant at $p < 0.05$ after a Bonferroni correction for multiple tests, and (iii) consensus breakpoints could be inferred from the results. Thus, "No recombination detected" does not preclude that intragenic recombination could be occurring within the sequence, but rather indicates that our stringent criteria have not identified such an event. While we define phages as either *Prochlorococcus* or *Synechococcus* phages depending on the original host of isolation, we note that many of the myoviruses cross-infect both genera (represented with a "Δ" where known, a "?" where unknown, and no symbol for isolates that do not cross-infect across genera). Consensus breakpoints are relative to nucleotide positions in *Thermosynechococcus psbA*.

Found at DOI: 10.1371/journal.pbio.0040234.st001 (29 KB XLS).

Table S2. Consensus Results from Four Tests for Intragenic Recombination within Gene Sequences in Our *psbD* Dataset

Details are as in Table S1.

Found at DOI: 10.1371/journal.pbio.0040234.st002 (28 KB XLS).

Table S3. PCR Conditions

Found at DOI: 10.1371/journal.pbio.0040234.st003 (38 KB DOC).

Table S4. PCR Primers

Found at DOI: 10.1371/journal.pbio.0040234.st004 (39 KB DOC).

Table S5. Likelihood Ratio Tests for Variable Evolutionary Rates among Branches

For both *psbA* and *psbD,* individual sequences exhibiting a signature for intragenic recombination (Tables S1 and S2) were excluded from analysis. Likelihood scores were obtained under a stationary HKY85 model combined with a gamma correction for among-sites rate variation. All model parameters, including nucleotide frequencies, were estimated by using maximum likelihood. Data analysis included all three codon positions. Models were employed as implemented in the baseml program of the PAML package [82]. Tree 1 was obtained by neighbor-joining analysis of LogDet distances estimated from all three codon positions. Tree 2 was obtained by neighbor-joining analysis of LogDet distances estimated from 1st and 2nd codon positions. For both genes, Tree 1 grouped lineages along lines of similarity in evolutionary rates and compositional biases, and Tree 2 did not.

Found at DOI: 10.1371/journal.pbio.0040234.st005 (36 KB DOC).

Table S6. Likelihood Ratio Tests for Nonstationary Frequencies among Lineages

H_0 denotes the null hypothesis of stationary nucleotide frequencies; this was modeled by specifying one set of nucleotide frequencies for all branches of the tree. H_1 denotes the alternative hypothesis of nonstationary nucleotide frequencies; this was modeled by assigning all branches of the tree topology to one of several independent sets of frequency parameters (six sets for *psbA* and five sets for *psbD*). Apart from nucleotide frequencies, H_0 and H_1 assumed a substitution process equivalent to an HKY85 model combined with a gamma model for among-sites rate variation. The transition/transversion ratio was assumed to be homogenous among branches. H_1 represents a user-defined version of the nonhomogenous models of Yang and Roberts [77]. All model parameters, including nucleotide frequencies, were estimated by using maximum likelihood. Data analysis included all three codon positions. Models were employed as implemented in the baseml program of the PAML package [82].
Tree 1 was obtained by neighbor-joining analysis of LogDet distances estimated from all three codon positions. Tree 2 was obtained by neighbor-joining analysis of LogDet distances estimated from 1st and 2nd codon positions. For both genes, Tree 1 grouped lineages along lines of similarity in evolutionary rates and compositional biases, and Tree 2 did not. User-defined sets of frequency parameters for H_1 were specified in the tree file (shown below) by using the "branch label" format described in the PAML manual. For both *psbA* and *psbD,* individual sequences exhibiting a signature for intragenic recombination (Tables S1 and S2) were excluded from analysis.

Found at DOI: 10.1371/journal.pbio.0040234.st006 (44 KB DOC).

Table S7. Likelihood-Based Statistical Comparison of Competing Evolutionary Hypotheses under a Model of Nonstationary Nucleotide Frequencies

P_{KH} denotes the *p*-value for the KH normal test of [83]. P_{SH} denotes the *p*-value for the SH test [84]. P_{RELL} denotes the RELL bootstrap proportion [83]. Note that although Tree 1 and Tree 2 were not selected independently of the data, neither was selected according to its likelihood score. For both genes, Tree 1 grouped lineages along lines of similarity in evolutionary rates and compositional biases, and Tree 2 did not. For both *psbA* and *psbD,* individual sequences exhibiting a signature for intragenic recombination (Tables S1 and S2) were excluded from analysis. Tree 1 was estimated by a neighbor-joining analysis of LogDet distances from all sites, and Tree 2 was estimated by a neighbor-joining analysis of LogDet distances based on only 1st and 2nd codon positions. Likelihood scores were obtained under nonstationary models of nucleotide frequencies (see Table S5 for additional model details).

Found at DOI: 10.1371/journal.pbio.0040234.st007 (46 KB DOC).

Accession Numbers

New sequences from cultured cyanobacteria and cyanophages are deposited in GenBank (http://www.ncbi.nlm.nih.gov/Genbank) under accession numbers DQ473647–DQ473719, whereas new environmental sequences are deposited under accession numbers DQ473720–DQ473847.

Acknowledgments

We thank M. Shmoish, R. Fu, and V. Quinlivan for technical assistance; M. Coleman, M. Osburne, J. Waldbauer, and V. Rich for valuable comments on the manuscript; A. Thompson for collecting field samples; and Z. Johnson, K. Armstrong, and B. Tidor for analysis and discussion of possible PsbA/PsbD interactions. We thank P. Weigele, W. Pope, G. Hatfull, and R. Hendrix for providing unpublished genome sequences (Syn5 and Syn9); and F. Chen for sharing his unpublished phage lytic cycle information (P60) with us.

Author contributions. MBS, DL, and SWC conceived and designed the experiments. MBS, DL, JAL, and LRT performed the experiments. MBS, DL, JAL, LRT, and JPB analyzed the data. MBS and DL wrote the paper, with significant contributions from all authors.

Funding. This research was supported by grants from the United States Department of Energy (DE-FG02-99ER62814 and DE-FG02-02ER63445), the National Science Foundation and the Gordon and Betty Moore Foundation to SWC, Massachusetts Institute of Technology's Undergraduate Research Opportunities Program funding to JAL, a National Institutes of Health predoctoral training grant in the biological sciences (GM07287–31) to LRT, and a National Sciences and Engineering Research Council (Canada) Discovery Grant (DG 298394) to JPB.

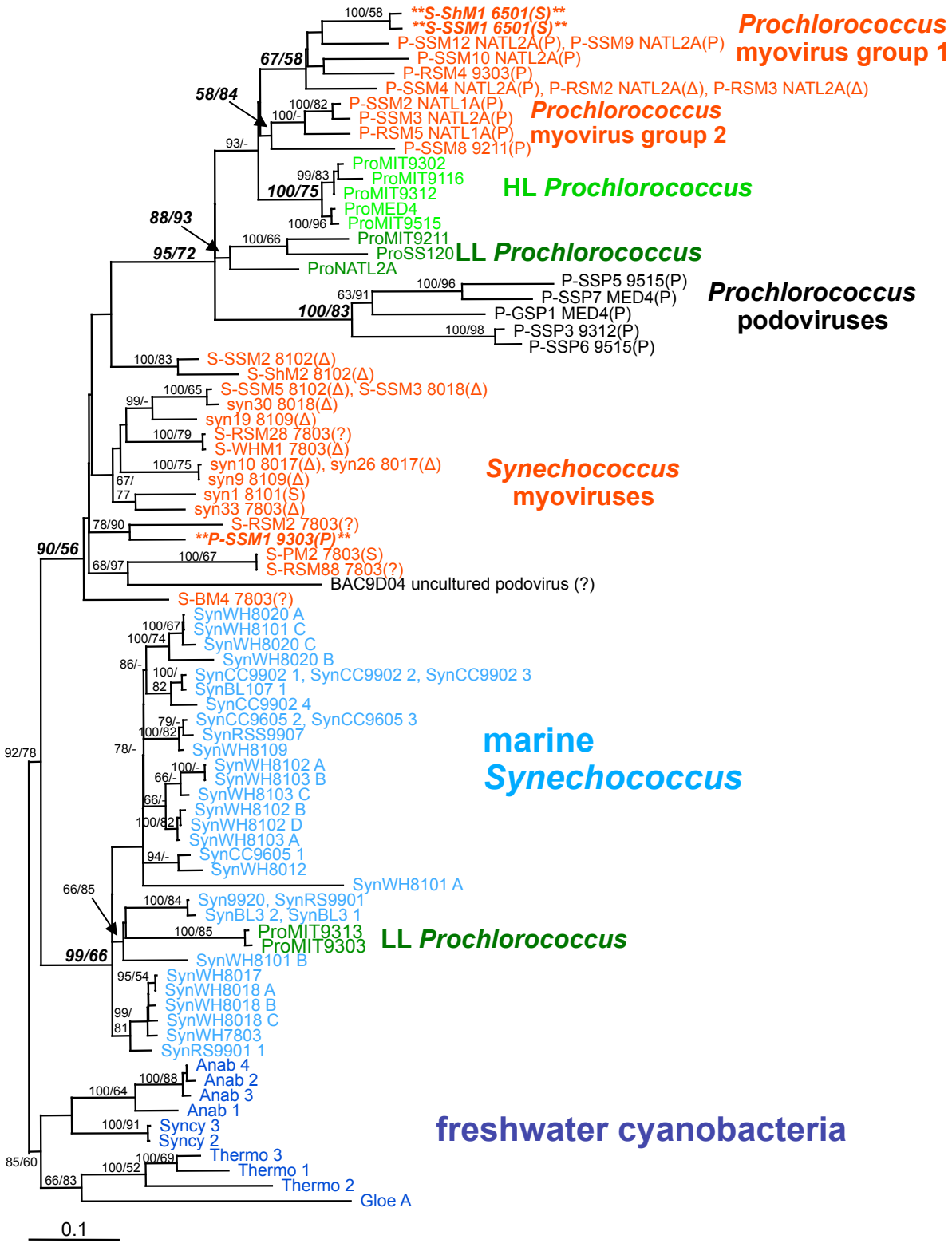
Competing interests. The authors have declared that no competing interests exist.

References

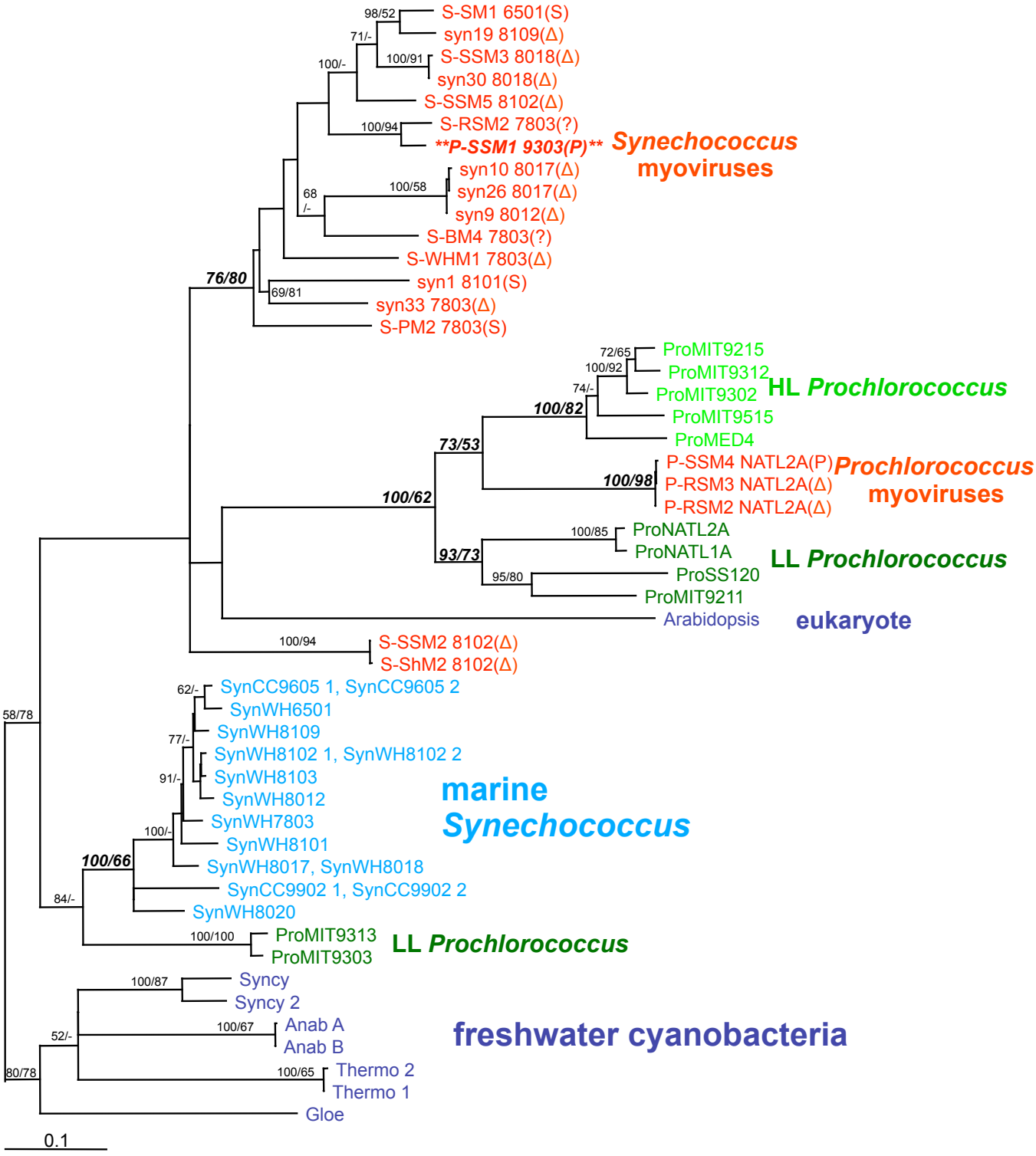
1. Waterbury JB, Watson SW, Valois FW, Franks DG (1986) Biological and ecological characterization of the marine unicellular cyanobacterium *Synechococcus*. Can Bull Fish Aquat Sci 214: 71–120.
2. Partensky F, Hess WR, Vaulot D (1999) *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. Microbiol Mol Biol Rev 63: 106–127.
3. Suttle CA, Chan AM (1994) Dynamics and distribution of cyanophages and their effects on marine *Synechococcus* spp. Appl Environ Microbiol 60: 3167–3174.
4. Waterbury JB, Valois FW (1993) Resistance to co-occurring phages enables marine *Synechococcus* communities to coexist with cyanophage abundant in seawater. Appl Environ Microbiol 59: 3393–3399.
5. Sullivan MB, Waterbury JB, Chisholm SW (2003) Cyanophages infecting the oceanic cyanobacterium *Prochlorococcus*. Nature 424: 1047–1051.
6. Lu J, Chen F, Hodson RE (2001) Distribution, isolation, host specificity, and diversity of cyanophages infecting marine *Synechococcus* spp. in river estuaries. Appl Environ Microbiol 67: 3285–3290.
7. Marston MF, Sallee JL (2003) Genetic diversity and temporal variation in the cyanophage community infecting marine *Synechococcus* species in Rhode Island's coastal waters. Appl Environ Microbiol 69: 4639–4647.
8. Muhling M, Fuller NJ, Millard A, Somerfield PJ, Marie D, et al. (2005) Genetic diversity of marine *Synechococcus* and co-occurring cyanophage communities: Evidence for viral control of phytoplankton. Environ Microbiol 7: 499–508.
9. Wilson WH, Joint IR, Carr NG, Mann NH (1993) Isolation and molecular characterization of five marine cyanophages propogated on *Synechococcus* sp. strain WH 7803. Appl Environ Microbiol 59: 3736–3743.
10. Suttle CA, Chan AM (1993) Marine cyanophages infecting oceanic and coastal strains of *Synechococcus*: Abundance, morphology, cross-infectivity and growth characteristics. Mar Ecol Prog Ser 92: 99–109.
11. Brussow H, Canchaya C, Hardt WD (2004) Phages and the evolution of bacterial pathogens: From genomic rearrangements to lysogenic conversion. Microbiol Mol Biol Rev 68: 560–602.
12. Faruque SM, Mekalanos JJ (2003) Pathogenicity islands and phages in *Vibrio cholerae* evolution. Trends Microbiol 11: 505–510.
13. Lindell D, Sullivan MB, Johnson ZI, Tolonen AC, Rohwer F, et al. (2004) Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. Proc Natl Acad Sci U S A 101: 11013–11018.
14. Canchaya C, Proux C, Fournous G, Bruttin A, Brussow H (2003) Prophage genomics. Microbiol Mol Biol Rev 67: 238–276.
15. Casjens S (2003) Prophages and bacterial genomics: What have we learned so far? Mol Microbiol 49: 277–300.
16. Forterre P (1999) Displacement of cellular proteins by functional analogues from plasmids or viruses could explain puzzling phylogenies of many DNA informational proteins. Mol Microbiol 33: 457–465.
17. Filee J, Forterre P, Laurent J (2003) The role played by viruses in the evolution of their hosts: A view based on informational protein phylogenies. Res Microbiol 154: 237–243.
18. Filee J, Forterre P, Sen-Lin T, Laurent J (2002) Evolution of DNA polymerase families: Evidences for multiple gene exchange between cellular and viral proteins. J Mol Evol 54: 763–773.
19. Hendrix RW (1999) Evolution: The long evolutionary reach of viruses. Curr Biol 9: R914–917.
20. Hendrix RW, Lawrence JG, Hatfull GF, Casjens S (2000) The origins and ongoing evolution of viruses. Trends Microbiol 8: 504–508.
21. Juhala RJ, Ford ME, Duda RL, Youlton A, Hatfull GF, et al. (2000) Genomic

sequences of bacteriophages HK97 and HK022: Pervasive genetic mosaicism in the lambdoid bacteriophages. J Mol Biol 299: 27–51.
22. Mann NH, Cook A, Millard A, Bailey S, Clokie M (2003) Bacterial photosynthesis genes in a virus. Nature 424: 741.
23. Botstein D (1980) A theory of modular evolution for bacteriophages. Ann New York Acad Sci 354: 484–491.
24. Hendrix RW, Smith MC, Burns RN, Ford ME, Hatfull GF (1999) Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. Proc Natl Acad Sci U S A 96: 2192–2197.
25. Silander OK, Weinreich DM, Wright KM, O'Keefe KJ, Rang CU, et al. (2005) Widespread genetic exchange among terrestrial bacteriophages. Proc Natl Acad Sci U S A 102: 19009–19014.
26. Filee J, Tetart F, Suttle CA, Krisch HM (2005) Marine T4-type bacteriophages, a ubiquitous component of the dark matter of the biosphere. Proc Natl Acad Sci U S A.
27. Breitbart M, Miyake JH, Rohwer F (2004) Global distribution of nearly identical phage-encoded DNA sequences. FEMS Microbiol Lett 236: 249–256.
28. Mann NH, Clokie MR, Millard A, Cook A, Wilson WH, et al. (2005) The genome of S-PM2, a "photosynthetic" T4-type bacteriophage that infects marine *Synechococcus*. J Bacteriol 187: 3188–3200.
29. Millard A, Clokie MR, Shub DA, Mann NH (2004) Genetic organization of the *psbAD* region in phages infecting marine *Synechococcus* strains. Proc Natl Acad Sci U S A 101: 11007–11012.
30. Sullivan MB, Coleman M, Weigele P, Rohwer F, Chisholm SW (2005) Three *Prochlorococcus* cyanophage genomes: Signature features and ecological interpretations. PLoS Biol 3: e144. DOI: 10.1371/journal.pbio.0030144
31. Lindell D, Jaffe JD, Johnson ZI, Church GM, Chisholm SW (2005) Photosynthesis genes in marine viruses yield proteins during host infection. Nature 438: 8689.
32. Clokie MRJ, Shan J, Bailey S, Jia Y, Krisch HM (2006) Transcription of a 'photosynthetic' T4-type phage during infection of a marine cyanobacterium. Environ Microbiol 8: 827–835.
33. Benson R, Martin E (1981) Effects of photosynthetic inhibitors and light-dark regimes on the replication of cyanophage SM-2. Arch Microbiol 129: 165–167.
34. Adir N, Zer H, Schochat S, Ohad I (2003) Photoinhibition—A historical perspective. Photosynth Res 76: 343–370.
35. Paul JH, Sullivan MB (2005) Marine phage genomics: What have we learned? Curr Opin Biotechnol 16: 299–307.
36. Chen F, Lu J (2002) Genomic sequence and evolution of marine cyanophage P60: A new insight on lytic and lysogenic phages. Appl Environ Microbiol 68: 2589–2594.
37. Zeidner G, Bielawski JP, Shmoish M, Scanlan DJ, Sabehi G, et al. (2005) Potential photosynthesis gene recombination between *Prochlorococcus* and *Synechococcus* via viral intermediates. Environ Microbiol 7: 1505–1513.
38. Sherman LA (1976) Infection of *Synechococcus cedrorum* by the cyanophage AS-1M. III. Cellular metabolism and phage development. Virology 71: 199–206.
39. Wilson WH, Carr NG, Mann NH (1996) The effect of phosphate status on the kinetics of cyanophage infection in the oceanic cyanobacterium *Synechococcus* sp. WH7803. J Phycol 32: 506–516.
40. Levin BR, Lenski RE (1983) Coevolution in bacteria and their viruses and plasmids. In: Futuyma DJ, Slatkin M, editors. Coevolution. Sunderland (Massachusetts): Sinauer. pp. 99–127.
41. Wang IN, Dykhuizen DE, Slobodkin LB (1996) The evolution of phage lysis timing. Evol Ecol 10: 545–558.
42. Abedon ST (1989) Selection for bacteriophage latent period length by bacterial density: A theoretical examination. Microb Ecol 18: 79–88.
43. Abedon ST, Herschler TD, Stopar D (2001) Bacteriophage latent-period evolution as a response to resource availability. Appl Environ Microbiol 67: 4233–4241.
44. Abedon ST, Hyman P, Thomas C (2003) Experimental examination of bacteriophage latent-period evolution as a response to bacterial availability. Appl Environ Microbiol 69: 7499–7506.
45. Moore LR, Rocap G, Chisholm SW (1998) Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. Nature 393: 464–467.
46. Lindell D, Penno S, Al-Qutob M, David E, Rivlin T, et al. (2005) Expression of the nitrogen stress response gene *ntcA* reveals nitrogen-sufficient *Synechococcus* populations in the oligotrophic northern Red Sea. Limnol and Oceanogr 50: 1932–1944.
47. Palenik B (2001) Chromatic adaptation in marine *Synechococcus* strains. Appl Environ Microbiol 67: 991–994.
48. Rocap G, Distel D, Waterbury JB, Chisholm SW (2002) Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. Appl Environ Microbiol 68: 1180–1191.
49. Fuller NJ, Marie D, Partensky F, Vaulot D, Post AF, et al. (2003) Clade-specific 16S ribosomal DNA oligonucleotides reveal the predominance of a single marine *Synechococcus* clade throughout a stratified water column in the Red Sea. Appl Environ Microbiol 69: 2430–2443.
50. Golden SS, Brusslan J, Haselkorn R (1986) Expression of a family of *psbA* genes encoding a photosystem II polypeptide in the cyanobacterium *Anacystis nidulans* R2. Embo J 5: 2789–2798.
51. Sicora CI, Appleton SE, Brown CM, Chung J, Chandler J, et al. (2006) Cyanobacterial *psbA* families in Anabaena and Synechocystis encode trace, constitutive and UVB-induced D1 isoforms. Biochim Biophys Acta 1757: 47–56.
52. Clarke AK, Soitamo A, Gustafsson P, Oquist G (1993) Rapid interchange between two distinct forms of cyanobacterial photosystem II reaction-center protein D1 in response to photoinhibition. Proc Natl Acad Sci U S A 90: 9973–9977.
53. Hess WR, Weihe A, Loiseaux-de Goer S, Partensky F, Vaulot D (1995) Characterization of the single *psbA* gene of *Prochlorococcus marinus* CCMP1375 (Prochlorophyta). Plant Mol Biol 27: 1189–1196.
54. Calendar R (1988) The bacteriophages. New York: Plenum.
55. Ackermann HW, DuBow MS (1987) Viruses of prokaryotes, Volume 1. General properties of bacteriophages. Boca Raton (Florida): CRC Press.
56. Karl DM (1999) A sea of change: Biogeochemical variability in the North Pacific Subtropical Gyre. Ecosystems 2: 181–214.
57. Zinser ER, Coe A, Johnson ZI, Martiny AC, Fuller NJ, et al. (2006) *Prochlorococcus* ecotype abundances in the North Atlantic Ocean as revealed by an improved quantitative PCR method. Appl Environ Microbiol 72: 723–732.
58. Dufresne A, Salanoubat M, Partensky F, Artiguenave F, Axmann IM, et al. (2003) Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. Proc Natl Acad Sci U S A 100: 10020–10025.
59. Palenik B, Brahamsha B, McCarren J, Waterbury J, Allen E, et al. (2003) The genome of a motile marine *Synechococcus*. Nature 424: 1037–1041.
60. Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, et al. (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature 424: 1042–1047.
61. Zeidner G, Preston CM, Delong EF, Massana R, Post AF, et al. (2003) Molecular diversity among marine picophytoplankton as revealed by *psbA* analyses. Environ Microbiol 5: 212–216.
62. Thompson JR, Marcelino LA, Polz MF (2002) Heteroduplexes in mixed-template amplifications: Formation, consequence and elimination by 'reconditioning PCR.' Nucleic Acids Res 30: 2083–2088.
63. Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, et al. (2006) Genomic islands and the ecology and evolution of *Prochlorococcus*. Science 311: 1768–1770.
64. Hess WR, Rocap G, Ting CS, Larimer FW, Stilwagen S, et al. (2001) The photosynthetic apparatus of *Prochlorococcus*: Insights through comparative genomics. Photosynth Res 70: 53–71.
65. Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF (2005) PCR-induced sequence artifacts and bias: Insights from comparison of two 16S rRNA clone libraries constructed from the same sample. Appl Environ Microbiol 71: 8966–8969.
66. Saitou N, Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406–425.
67. Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22: 160–174.
68. Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J Mol Evol 39: 306–314.
69. Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: Maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18: 502–504.
70. Phillips MJ, Delsuc F, Penny D (2004) Genome-scale phylogeny and the detection of systematic biases. Mol Biol Evol 21: 1455–1458.
71. Kennedy M, Holland BR, Gray RD, Spencer HG (2005) Untangling long branches: Identifying conflicting phylogenetic signals using spectral analysis, neighbor-net, and consensus networks. Syst Biol 54: 620–633.
72. Bryant D, Moulton V (2004) Neighbor-net: An agglomerative method for the construction of phylogenetic networks. Mol Biol Evol 21: 255–265.
73. Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23: 254–267.
74. Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 27: 401–410.
75. Lockhart PJ, Steel M, Hendy MD, Penny D (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. Mol Biol Evol 11: 605–612.
76. Schierup MH, Hein J (2000) Consequences of recombination on traditional phylogenetic analysis. Genetics 156: 879–891.
77. Yang Z, Roberts D (1995) On the use of nucleic acid sequences to infer early branchings in the tree of life. Mol Biol Evol 12: 451–458.
78. Padidam M, Sawyer S, Fauquet CM (1999) Possible emergence of new geminiviruses by frequent recombination. Virology 265: 218–225.
79. Maynard Smith J (1992) Analyzing the mosaic structure of genes. J Mol Evol 34: 126–129.
80. Posada D, Crandall KA (2001) Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. Proc Natl Acad Sci U S A 98: 13757–13762.
81. Martin D, Rybicki E (2000) RDP: Detection of recombination amongst aligned sequences. Bioinformatics 16: 562–563.
82. Yang Z (1997) PAML: A program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13: 555–556.
83. Kishino H, Hasegawa M (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J Mol Evol 29: 170–179.
84. Shimodaira H, Hasegawa M (1999) Multiple comparisons of log-likelihoods and combining nonnested models with applications to phylogenetic inference. Mol Biol Evol 16: 1114–1116.

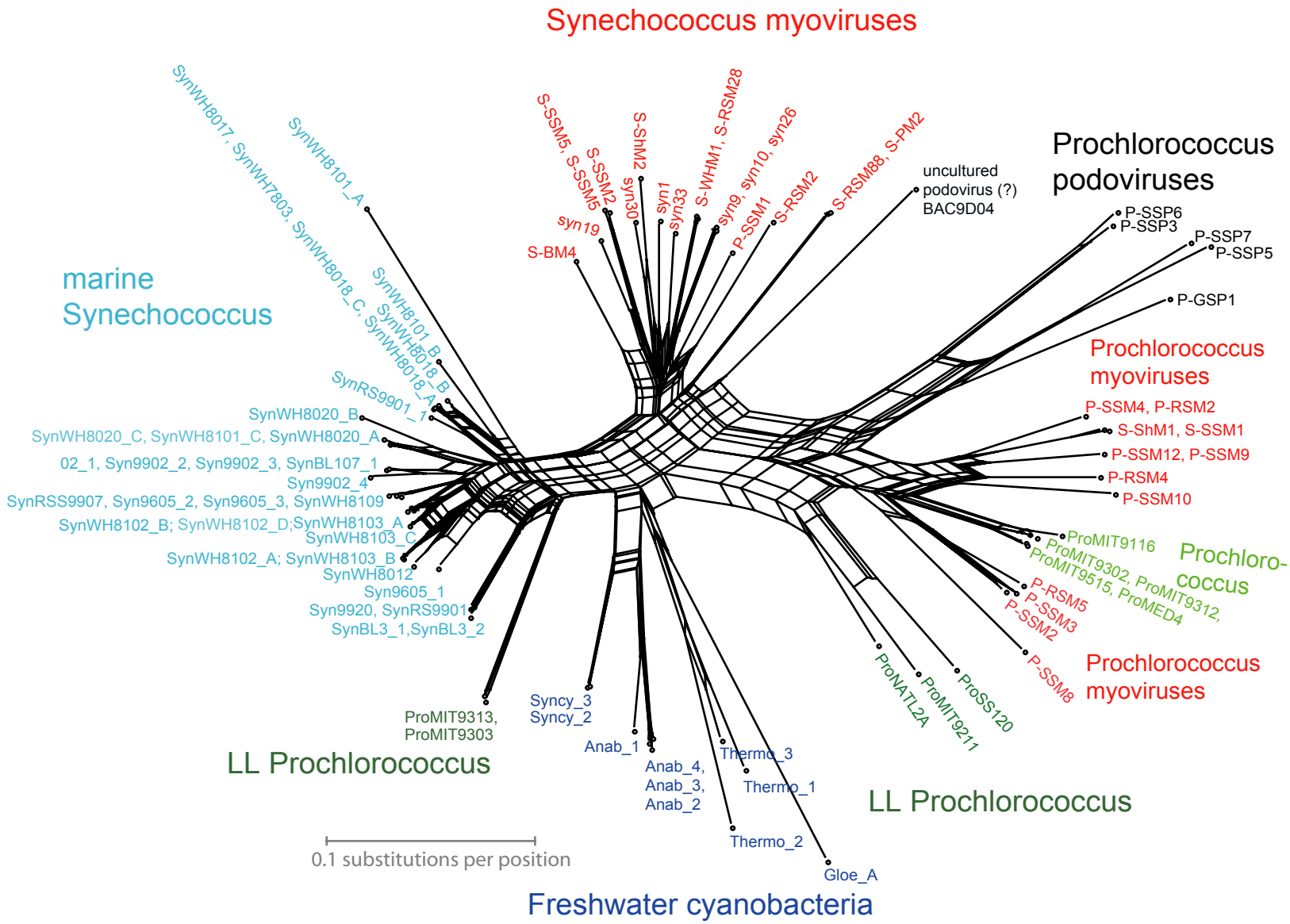
Suppl. Figure 1



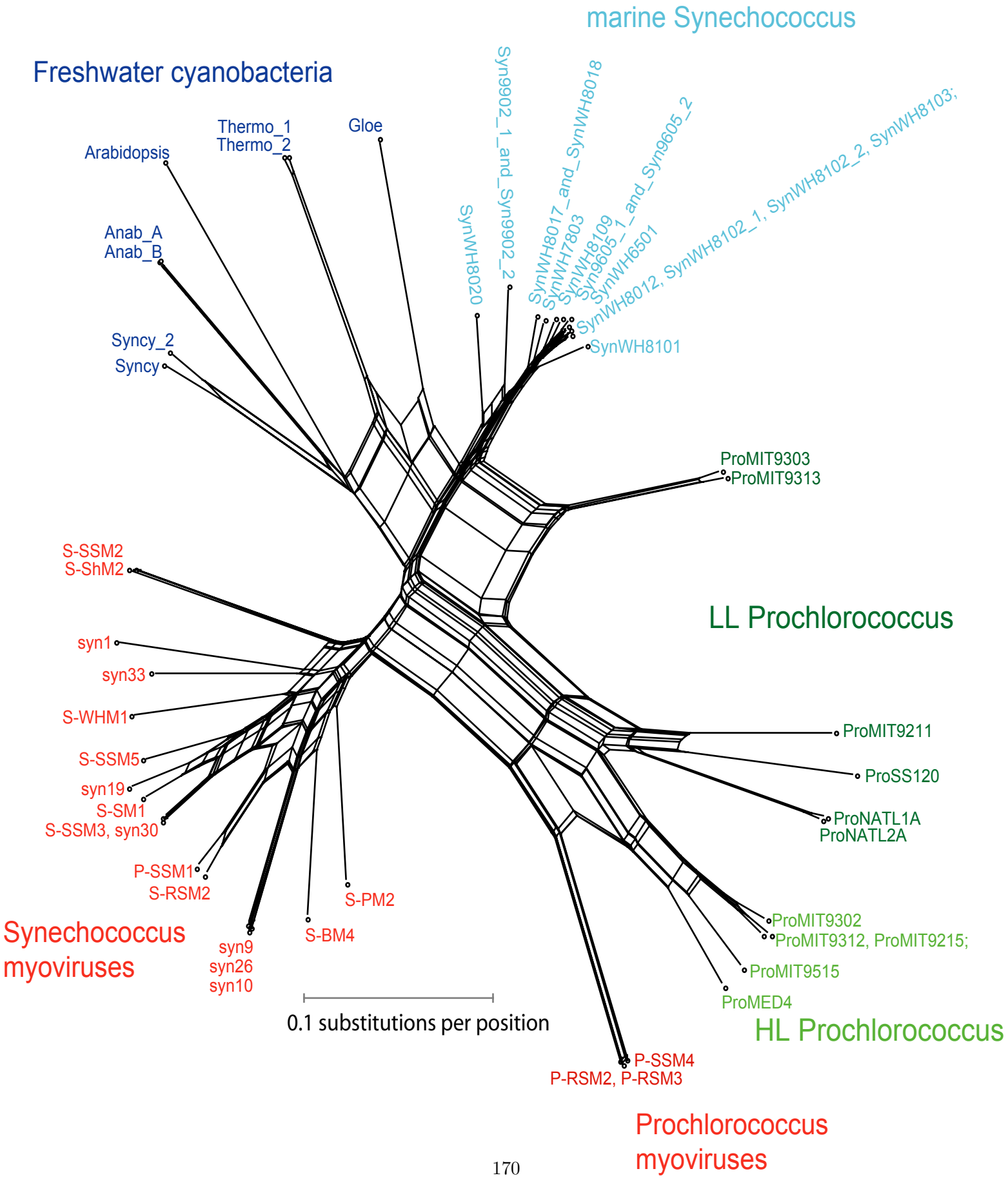
Suppl. Figure 2



Supplementary Figure 3



Supplementary Figure 4



Suppl. Table 1: Consensus results from 4 tests for intragenic recombination within gene sequences in our *psbA* dataset. The four tests included (1) RDP, (2) GeneConv, (3) MaxChi, (4) Chimaera (as described in methods), and recombination was only considered 'detected' when the following criteria were satisfied: (i) similar regions were detected by 2 or more methods, (ii) all such regions were significant at P<0.05 after a Bonferroni correction for multiple tests, (iii) consensus breakpoints could be inferred from the results. Thus, "No recombination detected" does not preclude that intragenic recombination could be occurring within the sequence, but rather indicates our stringent criteria have not identified such an event. While we define phages as either *Prochlorococcus* or *Synechococcus* phages depending on the original host of isolation, we note that many of the myoviruses cross infect both genera (represented with a 'Δ' where known, a '?' where unknown, and no symbol for isolates that do not cross-infect across genera). Concensus breakpoints relative to nucleotide positions in *Thermosynechococcus psbA*.

	No.	Recipient Taxon	Recombination detected?	Method	Consensus breakpoints, start…end	Donor of foreign sequence
Pro myoviruses (plus S-ShM1, S-SSM1*)	**1**	P-RSM4_9303	No			
	2	P-SSM12_NATL2A, P-SSM9_NATL2A	Yes	1,2	591..741	Unknown
	3	P-SSM4_NATL2A, P-RSM2_NATL2A Δ, P-RSM3_NATL2A Δ	No			
	4	S-ShM1_6501	No			
	5	S-SSM1_6501	No			
	6	P-SSM10_NATL2A	No			
HL Prochloro-coccus hosts	**7**	MED4	No			
	8	MIT9515	No			
	9	MIT9302	No			
	10	MIT9312	No			
	11	MIT9116	Yes	1,2	689..714	Unknown
Pro myoviruses	**12**	P-SSM2_NATL1A	No			
	13	P-SSM3_NATL2A	No			
	14	P-RSM5_NATL1A	No			
	15	P-SSM8_9211	No			
LL Prochloro-coccus hosts	**16**	MIT9211	No			
	17	SS120	No			
	18	NATL2A	No			
Synechococcus myoviruses (plus P-SSM1*)	**19**	S-SSM2_8102 Δ	Yes	2,3	253..561	S-ShM2_8102
				1,2,3,4	565..668	P-SSP7_MED4(**1,3,4**) or P-SSP3_9312(**2**)
	20	S-ShM2_8102 Δ	Yes	3,4	669..789	Unknown & MED4
	21	syn1_8101	No			
	22	syn33_8018 Δ	Yes	2,3	459..567	syn10_8017 or syn9_8109
				2.3	655..828	Unknown
				1,2	793..959	S-ShM2_8102
				1,2,3,4	864..964	Syn9902_1 or 2 or 3(**1**), Unknown(**2**), WH8109(**3**), WH8017(**4**)
	23	S-SSM5_8102 Δ, S-SSM3_8018 Δ	No			
	24	syn30_8018 Δ	No			
	25	syn19_8109 Δ	No			
	26	S-BM4_7803 ?	No			
	27	S-RSM28_7803 ?	No			
	28	S-WHM1_7803 Δ	No			
	29	syn10_8017 Δ, syn26_8017 Δ	No			
	30	syn9_8109 Δ	No			
	31	S-RSM2_7803 ?	Yes	1,2,3,4	549..825	P-SSP7_MED4(**1,2**) or S-SSM2_8102(**3,4**)
	32	P-SSM1_9303	No			
	33	S-PM2_7803	No			
	34	S-RSM88_7803	No			
	35	BAC9D04	No			
Synechococcus hosts	**36**	WH8102_A	Yes	3,4	340..391	Unknown
				3,4	450..528	Syn9605_1
	37	WH8103_B	Yes	3,4	340..391	Unknown
				3,4	450..528	Syn9605_1
	38	WH8103_C	Yes	3,4	448..631	Unknown
				1,3	405..548	Unknown
				1,2	450..528	Unknown
	39	WH8102_B	No			
	40	WH8102_D	No			
	41	WH8103_A	No			
	42	Syn9902_1, Syn9902_2, Syn9902_3	No			
	43	SynBL107_1	Yes	2,3	830..951	S-ShM2_8102 or Syn9605_psbA2
	44	Syn9902_4	No			
	45	WH8020_A	Yes	1,2,3	961..976	MED4 or P-SSP6_9515 or Unknown
	46	WH8101_C	Yes	1,2,3	961..977	MED4 or P-SSP6_9515 or Unknown
	47	WH8020_C	No			
	48	WH8020_B	No			
	49	Syn9605_2, Syn9605_3	No			
	50	SynRSS9907	No			
	51	WH8109	No			
	52	Syn9605_1	Yes	3,4	867..977	Unknown
	53	WH8012	No			
	54	WH8017	No			
	55	WH8018_A	No			
	56	WH8018_C	No			
	57	WH7803	No			
	58	WH8018_B	No			
	59	SynRS9901_1	No			
	60	Syn9920, SynRS9901	Yes	2,3	642..846	WH8018_B or WH8017
	61	SynBL3_2, SynBL3_1	Yes	2,3	642..847	WH8018_B or WH8018
	62	MIT9313	No			
	63	MIT9303	No			
	64	WH8101_B	No			
	65	WH8101_A	Yes	3,4	869..929	Unknown
Pro-chloro-coccus podo-viruses	**66**	PSSP5_9515	No			
	67	PSSP7_MED4	No			
	68	PGSP1_MED4	No			
	69	PSSP3_9312	No			
	70	PSSP6_9515	No			

* Prochlorococcus myovirus P-SSM1 and Synechococcus myoviruses S-ShM1 and S-SSM1 have *psbA* gene sequences that do not cluster as predicted (see text)

Suppl. Table 2: Consensus results from 4 tests for intragenic recombination within gene sequences in our *psbD* dataset. The four tests included (1) RDP, (2) GeneConv, (3) MaxChi, (4) Chimaera (as described in methods), and recombination was only considered 'detected' when the following criteria were satisfied: (i) similar regions were detected by 2 or more methods, (ii) all such regions were significant at P<0.05 after a Bonferroni correction for multiple tests, (iii) consensus breakpoints could be inferred from the results. Thus, "No recombination detected" does not preclude that intragenic recombination could be occurring within the sequence, but rather indicates our stringent criteria have not identified such an event. While we define phages as either *Prochlorococcus* or *Synechococcus* phages depending on the original host of isolation, we note that many of the myoviruses cross infect both genera (represented with a 'Δ' where known, a '?' where unknown, and no symbol for isolates that do not cross-infect across genera). Concensus breakpoints relative to nucleotide positions in *Thermosynechococcus psbD*.

	No.	Recipient Taxon	Recombination detected?	Method	Consensus breakpoints, start…end	Donor of foreign sequence
Prochlorococcus hosts	**1**	MIT9211	No			
	2	SS120	No			
	3	NATL2A	No			
	4	NATL1A	No			
	5	MIT9215	No			
	6	MIT9312	No			
	7	MIT9302	No			
	8	MIT9515	No			
	9	MED4	No			
Pro Myo viruses	**10**	P-SSM4_NATL2A	Yes	2, 3	714..822	Unknown
	11	P-RSM2_NATL2A Δ	Yes	2, 3	714..823	Unknown
	12	P-RSM3_NATL2A Δ	Yes	2, 3	714..824	Unknown
Synechococcus myoviruses (plus P-SSM1*)	**13**	S-SSM2_8102 Δ	Yes	1,2,3,4	192..369	S-BM4 or P-SSM1_9303
	14	S-ShM2_8102 Δ	Yes	1,2,3,5	193..370	S-BM4 or P-SSM1_9303
	15	syn10_8017 Δ	Yes	1,3	183..255	Unknown
				1,3	624..871	Unknown
	16	syn26_8017 Δ	Yes	1,3	183..255	Unknown & MIT 9303(**3**)
				1,3	624..871	Unknown
	17	syn9_8012 Δ	Yes	1,3	183..255	Unknown
				1,3	624..871	Unknown
	18	S-BM4_7803 ?	Yes	1,2,4	183..255	Unknown
				1,3,4	372..432	Unknown & S-SSM3_8018 (**3**)
				1,2,3,4	624..851	Unknown
	19	S-RSM2_7803 ?	No			
	20	P-SSM1_9303	Yes	1,2,4	183..253	Unknown
				1,3,4	496..606	Unknown & WH8017 & WH8101
				1,2,4	621..871	Unknown
	21	S-SM1_6501	Yes	3,4	189..300	syn1_8101 & Unknown
				2,3,4	639..837	P-SSM1_9303
	22	syn19_8109 Δ	Yes	1,3,4	189..300	Unknown
				2,3	639..837	S-RSM2_7808
				1,4	709..786	P-SSM1_9303
	23	S-SSM3_8018 Δ	Yes	1,2,3,4	216..252	S-BM4 & P-SSM1_9303 (**2,3**) & S-RSM2_7803
				3,4	357..465	Unknown & S-WHM1_7803
	24	syn30_8018 Δ	Yes	1,2,3,4	216..252	S-BM4 & P-SSM1_9303 (**2,3**) & S-RSM2_7803
				3,4	357..465	Unknown & syn1_8101
	25	S-SSM5_8102 Δ	Yes	1,2,3,4	317..453	Unknown
				1,3,4	459..618	Unknown & S-SSM3_8018
				1,4	483..541	syn33_7803
				2,3,4	639..837	P-SSM1_9303
	26	S-WHM1_7803 Δ	Yes	2,4	369..483	syn1_8101
				2,4	468..579	Unknown & WH8017
	27	syn1_8101	Yes	1,4	486..634	Unknown
	28	syn33_7803 Δ	Yes	1,2,3,4	601..864	S-SSM2_8102 & S-ShM2_8102
	29	S-PM2_7803	No			
Synechococcus hosts	**30**	Syn9605_1 , Syn9605_2	No			
	31	WH6501	Yes	3,4	628..702	Unknown
	32	WH8109	No			
	33	WH8102_1 , WH8102_2	No			
	34	WH8103	No			
	35	WH8012	No			
	36	WH7803	Yes	3,4	567..639	S-SSM5 & Unknown
	37	WH8017, WH8018	Yes	3,4	624..701	Unknown
	38	WH8101	No			
	39	Syn9902_1, Syn9902_2	No			
	40	WH8020	No			
	41	MIT9313	No			
	42	MIT9303	No			

* Prochlorococcus myovirus P-SSM1 has both a *psbA* and *psbD* gene sequence that clusters with Synechococcus myoviruses (see text)

Supplementary Table 3. PCR conditions

PCR Protocol	Forward Primer	Reverse Primer	Primer Conc. (μ*M*)	dNTP Conc. (μ*M*)	$MgCl_2$ Conc. (m*M*)	Units of *Taq*	Reaction Volume (μL)	Initial Denaturation	# of Cycles	Cycled Denaturation	Cycled Annealing	Cycled Extension	Final Extension
A	*psbA*-F	*psbA*-R	0.25	200	2.5	2.0	20.0	94°C, 5 min	35	94°C, 1 min	52°C, 1 min	72°C, 1.5 min	72°C, 10 min
B	Pro-*psbA*-F	Pro-*psbA*-R	0.25	80	5.0	2.5	50.0	92°C, 4 min	35	92°C, 1 min	50°C, 1 min	68°C, 1 min	68°C, 10 min
C	*psbD*-26F	*psbD*-308R	1.00	200	1.5	1.0	20.0	94°C, 5 min	35	94°C, 1 min	51°C, 1 min	72°C, 1 min	72°C, 10 min
D	*psbD*-54F	*psbD*-308R	1.00	200	1.5	1.0	20.0	94°C, 5 min	35	94°C, 1 min	51°C, 1 min	72°C, 1 min	72°C, 10 min
E	g20-F	g20-R	1.25	250	1.5	1.0	20.0	94°C, 3 min	35	94°C, 15 s	35°C, 1 min	73°C, 1 min	73°C, 10 min
F	DNApol-90F	DNApol-355R	4	200	0.5 mM for lysates 1.5 mM for extracted DNA	1.0	20.0	94°C, 4 min	35	94°C, 1 min	35°C, 1 min	72°C, 1 min	72°C, 10 min

Supplementary Table 4. PCR primers

Short Name	Full Name	Sequence	Source	Purpose
psbA-F	58-VDIDGIREP-66	5′-GTNGAYATHGAYGGNATHMGNGARCC-3′	Zeidner *et al*. 2003	*psbA* screening
psbA-R	331-MHERNAHNFP-340	5′-GGRAARTTRTGNGCRTTNCKYTCRTGCAT-3′	Zeidner *et al*. 2003	*psbA* screening
Pro-*psbA*-F	Pro-*psbA*-1F	5’-AACATCATYTCWGGTGCWGT-3’	Z. Johnson	*psbA* screening
Pro-*psbA*-R	Pro-*psbA*-1R	5’-TCGTGCATTACTTCCATACC-3’	Z. Johnson	*psbA* screening
psbD-26F	*psbD*-26Fa *psbD*-26Fb	5'-TTYGTNTTYRTNGGNTGGAGYGG-3' 5'-TTYGTNTTYRTNGGNTGGTCNGG-3'	J. A. Lee and D. Lindell	*psbD* screening
psbD-54F	*psbD*-54Fa *psbD*-54Fb	5'-GTNACNAGYTGGTAYACNCAYGG-3' 5'-GTNACNTCNTGGTAYACNCAYGG-3'	J. A. Lee and D. Lindell	*psbD* screening
psbD-308R	*psbD*-308Ra *psbD*-308Rb	5'-YTCYTGNGANACRAARTCRTANGC-3' 5'-YTCYTGRCTNACRAARTCRTANGC-3'	J. A. Lee and D. Lindell	*psbD* screening
g20-F	CPS1.1	5’-GTAGWATWTTYTAYATTGAYGTWGG-3’	M.B. Sullivan	*g20* screening
g20-R	CPS8.1	5’-ARTAYTTDCCDAYRWAWGGWTC-3’	M.B. Sullivan	*g20* screening
DNApol-F	DNApol-90Fa DNApol-90Fb	5’-GAYACIYTIRYIYTITCIMG-3’ 5’-GAYACIYTIRYIYTIAGYMG-3’	D. Lindell	*DNApol* screening
DNApol-R	DNApol-355Ra DNApol-355Rb	5’-GGIAYYTGIGCIARRTTIGG-3’ 5’-GGIAYRTTIGCIARRTTIGG-3’	D. Lindell	*DNApol* screening

Supplementary Table 5. Likelihood ratio tests (LRTs) for variable evolutionary rates among branches.

	H_0: clock		H_1: no clock		LRT		
	NP[1]	ℓ	NP[1]	ℓ	$2\Delta\ell$	df	*P* value
psbA							
Tree 1	63	-13611.73	125	-13370.58	482.3	62	$P < 0.0001$
Tree 2	63	-13845.70	125	-13448.08	795.2	62	$P < 0.0001$
psbD							
Tree 1	41	-10045.10	81	-9941.05	208.1	40	$P < 0.0001$
Tree 2	41	-10168.01	81	-9956.60	422.8	40	$P < 0.0001$

For both *psbA* and *psbD*, individual sequences exhibiting a signature for intragenic recombination (Suppl. Table 1, 2) were excluded from analysis. Likelihood scores were obtained under stationary HKY85 model combined with a gamma correction for among sites rate variation. All model parameters, including nucleotide frequencies, were estimated by using maximum likelihood. Data analysis included all three codon positions. Models were employed as implemented in the baseml program of the PAML package (Yang, 1997). Tree 1 was obtained by Neighbour-Joining analysis of LogDet distances estimated from all three codon positions. Tree 2 was obtained by Neighbour-Joining analysis of LogDet distances estimated from 1st and 2nd codon positions. For both genes, Tree 1 grouped lineages along lines of similarity in evolutionary rates and compositional biases, and Tree 2 did not.

[1] NP indicates the number of free branch length parameters [node times in the clock model] in the tree topology.

Supplementary Table 6. Likelihood ratio tests (LRTs) for non-stationary frequencies among lineages.

	H_0		H_1		LRT		
	NP[1]	ℓ	NP[1]	ℓ	$2\Delta\ell$	df	P value
psbA[2]							
Tree 1	3(1)	-10459.77	15(5)	-10274.38	370.8	12	$P < 0.0001$
Tree 2	3(1)	-10512.53	15(5)	-10332.88	359.3	12	$P < 0.0001$
psbD[3]							
Tree 1	3(1)	-9941.05	18(6)	-9785.28	311.5	15	$P < 0.0001$
Tree 2	3(1)	-9956.60	18(6)	-9808.98	295.2	15	$P < 0.0001$

H_0 denotes the null hypothesis of stationary nucleotide frequencies; this was modelled by specifying one set of nucleotide frequencies for all branches of the tree. H_1 denotes the alternative hypothesis of non-stationary nucleotide frequencies; this was modelled by assigning all branches of the tree topology to one of several independent sets of frequency parameters (6 sets for *psbA* and 5 sets for *psbD*). Apart from nucleotide frequencies, H_0 and H_1 assumed a substitution process equivalent to an HKY85 model combined with a gamma model for among sites rate variation. The transition to transversion ratio was assumed to be homogenous among branches. H_1 represents a user-defined version of the non-homogenous models of Yang and Roberts (1995). All model parameters, including nucleotide frequencies, were estimated by using maximum likelihood. Data analysis included all three codon positions. Models were employed as implemented in the baseml program of the PAML package (Yang, 1997).
Tree 1 was obtained by Neighbour-Joining analysis of LogDet distances estimated from all three codon positions. Tree 2 was obtained by Neighbour-Joining analysis of LogDet distances estimated from 1st and 2nd codon positions. For both genes, Tree 1 grouped lineages along lines of similarity in evolutionary rates and compositional biases, and Tree 2 did not. User defined sets of frequency parameters for H_1 were specified in the tree file (shown below) by using the "branch label" format described in the PAML manual. For both *psbA* and *psbD*, individual sequences exhibiting a signature for intragenic recombination (Suppl. Tables 1, 2) were excluded from analysis.

[1] NP indicates the number of free parameters for nucleotide frequencies. Number in parentheses indicates the number of independent sets (A+C+G+T) of user specified frequency parameters.

[2] *psbA* Tree 1:
((((Syncy_3,Syncy_2),(Anab_1,(Anab_3,(Anab_4,Anab_2)))),(Gloe_A,(Thermo_2,(Thermo_3,Thermo_1)))),(((WH8101_B #1,(((WH8103_A #1,(WH8102_B #1,WH8102_D #1)#1)#1,((WH8012 #1,((Syn9902_psbA1_Syn9902_psbA2_Syn9902_psbA3 #1,Syn9902_psbA4 #1)#1,(WH8020_C #1,WH8020_B #1)#1)#1)#1,(WH8109 #1,(Syn9605_psbA2_Syn9605_psbA3 #1,SynRSS9907 #1)#1)#1)#1)#1,(SynRS9901_1 #1,(WH7803 #1,(WH8018_C #1,(WH8018_B #1,(WH8017 #1,WH8018_A #1)#1)#1)#1)#1)#1),(MIT9313 #1,MIT9303 #1)#1)#1,((syn1_8101 #2,(PSSM1_9303 #2,((SBM4 #2,(syn19_8109 #2,(SSSM5_SSSM3_8018 #2,syn30_8018 #2)#2)#2)#2,((SRSM28_7803 #2,SWHM1_7803 #2)#2,(syn10_8017_Syn26_8017 #2,syn9_8109 #2)#2)#2)#2)#2)#2,((BAC9D04 #2,(SPM2_7803 #2,SRSM88_7803 #2)#2)#2,(((PGSP1_MED4 #3,(PSSP5_9515 #3,PSSP7_MED4 #3)#3)#3,(PSSP3_9312 #3,PSSP6_9515 #3)#3)#3,((NATL2A #3,(MIT9211 #3,SS120 #3)#3)#3,((PSSM8_9211 #3,(PRSM5_1A #3,(PSSM2_NATL1A #3,PSSM3_2A #3)#3)#3)#3,(((MED4 #3,MIT9515 #3)#3,(MIT9302 #3,MIT9312 #3)#3)#3,((PSSM4_2A_PRSM2_2A #3,(SShM1_6501 #3,SSSM1_6501 #3)#3)#3,(PSSM10_2A #3,PRSM4_9303 #3)#3)#3)#3)#3)#3)#3)#2)#2)#4);
[2] *psbA* Tree 2:
((((Syncy_3,Syncy_2),(Anab_1,(Anab_3,(Anab_4,Anab_2)))),(Gloe_A,(Thermo_2,(Thermo_3,Thermo_1)))),((WH8101_B #1,((((syn1_8101 #2,(PSSM1_9303 #2,(((syn19_8109 #2,(SSSM5_SSSM3_8018 #2,syn30_8018 #2)#2)#2,(SBM4 #2,(SRSM28_7803 #2,SWHM1_7803 #2)#2)#2)#2,(syn10_8017_Syn26_8017 #2,syn9_8109 #2)#2))#2)#2,(BAC9D04 #2,(SPM2_7803 #2,SRSM88_7803 #2)#2)#2)#2,((WH8102_D #1,(WH8102_B #1,WH8103_A #1)#1)#1,((Syn9902_psbA1_Syn9902_psbA2_Syn9902_psbA3 #1,(WH8020_C #1,(Syn9902_psbA4 #1,(WH8020_B #1,WH8012 #1)#1)#1)#1)#1,(SynRSS9907 #1,(Syn9605_psbA2_Syn9605_psbA3 #1,WH8109 #1)#1)#1)#1)#1,(SynRS9901_1 #1,(WH8018_B #1,(WH7803 #1,(WH8018_A #1,(WH8017 #1,WH8018_C #1)#1)#1)#1)#1)#1)#1,(((PGSP1_MED4 #3,(PSSP5_9515 #3,PSSP7_MED4 #3)#3)#3,(PSSP3_9312 #3,PSSP6_9515 #3)#3)#3,((MIT9313 #3,MIT9303 #3)#3,((MIT9211 #3,SS120 #3)#3,((NATL2A #3,((PSSM2_NATL1A #3,PSSM3_2A #3)#3,(PRSM5_1A #3,PSSM8_9211 #3)#3)#3)#3,(((MED4 #3,MIT9515 #3)#3,(MIT9302 #3,MIT9312 #3)#3)#3,(PSSM10_2A #3,(PRSM4_9303 #3,(PSSM4_2A_PRSM2_2A #3,(SShM1_6501 #3,SSSM1_6501 #3)#3)#3)#3)#3)#3)#3)#3)#3)#3)#4);

[3]*psbD* Tree 1:
((Gloe,(((Syncy,Syncy_2),(Arabidopsis,(Anab_A,Anab_B))),(Thermo_2,Thermo_1))),((((WH8101 #1,((Syn9605_1_Syn9605_2 #1,WH8109 #1)#1,(WH8012 #1,(WH8102_1_WH8102_2 #1,WH8103 #1)#1)#1)#1)#1,(Syn9902_1_Syn9902_2 #1,WH8020 #1)#1)#1,(MIT9313 #1,MIT9303 #1)#1)#1,((((SSSM2 #2,SShM2_8102 #2)#2,(syn1_8101 #2,Syn33_7803 #2)#2)#2,(SPM2_7803 #2,((SRSM2_7803 #2,(SBM4 #2,(syn9_8012 #2,(syn10_8017 #2,Syn26_8017 #2)#2)#2)#2)#2,(SSSM5_8102 #2,((SSM1_6501 #2,syn19_8109 #2)#2,(SSSM3_8018 #2,syn30_8018 #2)#2)#2)#2)#2)#2)#2,((MED4 #3,(MIT9515 #3,(MIT9302 #3,(MIT9215 #3,MIT9312 #3)#3)#3)#3)#3,((NATL2A #3,NATL1A #3)#3,(SS120 #3,MIT9211 #3)#3)#3)#3)#4)#5);
[3]*psbD* Tree 2:
((Gloe,((Syncy,Syncy_2),((Arabidopsis,(Anab_A,Anab_B)),(Thermo_2,Thermo_1)))),(((SPM2_7803 #2,(SBM4 #2,(((Syn33_7803 #2,(syn1_8101 #2,(SSSM2 #2,SShM2_8102 #2)#2)#2)#2,(SSSM3_8018 #2,syn30_8018 #2)#2)#2,((SRSM2_7803 #2,(syn9_8012 #2,(syn10_8017 #2,Syn26_8017 #2)#2)#2)#2,(SSSM5_8102 #2,(SSM1_6501 #2,syn19_8109 #2)#2)#2)#2)#2)#2)#2,(WH8020 #1,(WH8101 #1,((Syn9605_1_Syn9605_2 #1,(WH8109 #1,Syn9902_1_Syn9902_2 #1)#1)#1,(WH8012 #1,(WH8102_1_WH8102_2 #1,WH8103 #1)#1)#1)#1)#1)#4,((MIT9313 #3,MIT9303 #3)#3,((MED4 #3,(MIT9515 #3,(MIT9302 #3,(MIT9215 #3,MIT9312 #3)#3)#3)#3)#3,((NATL2A #3,NATL1A #3)#3,(SS120 #3,MIT9211 #3)#3)#3)#3)#3)#5);

Supplementary Table 7. Likelihood-based statistical comparison of competing evolutionary hypotheses under a model of non-stationary nucleotide frequencies.

Data partition	ℓ	$\Delta\ell$	P_{KH}	P_{SH}	P_{RELL}
psbA, All sites					
Tree 1	-13147.50	NA	NA	NA	0.75
Tree 2	-13182.14	-34.6	0.245	0.252	0.25
psbA, 1st and 2nd codon positions					
Tree 1	-4117.33	-90.4	0.004	0.004	0.003
Tree 2	-4026.88	NA	NA	NA	0.997
psbD, All sites					
Tree 1	-9785.28	NA	NA	NA	0.76
Tree 2	-9808.98	-23.7	0.239	0.239	0.24
psbA, 1st and 2nd codon positions					
Tree 1	-3020.89	-45.770	0.012	0.016	0.009
Tree 2	-2975.12	NA	NA	NA	0.991

P_{KH} denotes the *P*-value for the KH normal test of Kishino and Hasegawa (1989). P_{SH} denotes the *P*-value for the SH test (Shimodaira and Hasegawa, 1999). P_{RELL} denotes the RELL bootstrap proportion (Kishino and Hasegawa, 1989). Note that although tree 1 and tree 2 were not selected independently of the data, neither was selected according to its likelihood score. For both genes, Tree 1grouped lineages along lines of similarity in evolutionary rates and compositional biases, and Tree 2 did not. For both *psbA* and *psbD*, individual sequences exhibiting a signature for intragenic recombination (Suppl. Tables 1, 2) were excluded from analysis. Tree 1 was estimated by a Neighbour-Joining analysis of LogDet distances from all sites, and Tree 2 was estimated by a Neighbour-Joining analysis of LogDet distances based on only 1st and 2nd codon positions. Likelihood scores were obtained under non-stationary models of nucleotide frequencies (see Suppl. Table 5 for additional model details).

APPENDIX E

Exploring the vast diversity of marine viruses (Breitbart et al., *Oceanography*, 2007)

Exploring the Vast Diversity of Marine Viruses

BY MYA BREITBART, LUKE R. THOMPSON, CURTIS A. SUTTLE, AND MATTHEW B. SULLIVAN

At abundances routinely greater than 10 million particles per milliliter, viruses are the most numerous biological entities[1] in the oceans. To put the sheer abundance of marine viruses in context, we note that they contain more carbon than 75 million blue whales and, if such viruses were joined end-to-end, they would stretch further than the nearest 60 galaxies (Suttle, 2005). While marine viruses were first described by Spencer (1955), they were largely ignored for three decades because of the relatively low abundances inferred using culture-based assays. However, since Bergh et al. (1989) recognized their numeric importance, they have been considered at least as abundant as marine microbes, and scientists have been characterizing them and trying to determine the extent of marine viral diversity. Extensive efforts have focused on understanding the role of viruses in horizontal gene transfer and microbial mortality, and on the consequent impacts on microbial abundance, diversity, and community structure.

Here, we review advances in understanding viral diversity and genome evolution, and discuss potentially fruitful areas for future research. Our emerging view of the virosphere, inferred from gigabases of sequence data ground truthed by model systems in culture, is one of immense but finely tuned genetic diversity, where viruses have seemingly endless genetic potential, yet clearly are maintaining key genetic elements to propagate their extraordinary success.

One focus area is the diversity of marine viruses and marine viral communities. Although viruses might defy traditional species concepts, it is clear that viral genetic diversity is extremely high. Mathematical modeling based on viral metagenomic data predicts that there are hundreds of thousands of viral genotypes in the world's ocean (Angly et al., 2006). This may not be surprising given that marine microbial prokaryotic and eukaryotic diversity is also enormous (e.g., Irigoien et al., 2004; Witman et al., 2004; Thompson et al., 2005; Worden, 2006), and there are likely to be multiple host-specific viruses infecting each marine organism (Moebus, 1991; Moebus, 1992; Waterbury and Valois, 1993; Wilson et al., 1993; Wichels et al., 1998; Sullivan et al., 2003). The diversity of marine viral morphologies ranges from a variety of icosahedral tailed phages (Figure 1) (Moebus, 1991; Moebus, 1992; Waterbury and

> ...65–95% of marine viral metagenomic sequences are not similar to previously described sequences, as opposed to ~ 10% for cellular metagenomic surveys.

MYA BREITBART *(mya@marine.usf.edu) is Assistant Professor, College of Marine Science, University of South Florida, St. Petersburg, FL, USA.* LUKE R. THOMPSON *is Ph.D. Candidate, Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA.* CURTIS A. SUTTLE *is Professor, Departments of Earth & Ocean Sciences, Botany, and Microbiology & Immunology, University of British Columbia, Vancouver, Canada.* MATTHEW B. SULLIVAN *is Postdoctoral Associate, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA.*

[1] Viruses themselves are nonmetabolic (outside of the infection process) and lack the standard genetic marker (ribosomal RNA) that allows routine genetic comparison of known and unknown life forms using the "Tree of Life," so they are often not considered "alive." The term "biological entities" is used to allow classification of viruses with other life forms.

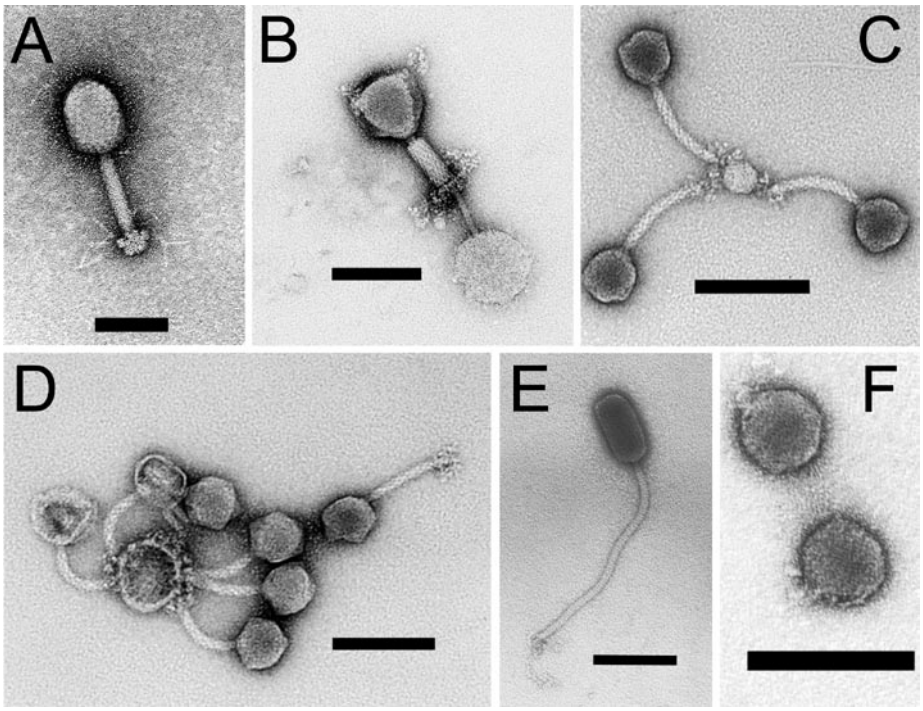


Figure 1. Electron micrographs of representative ocean cyanobacterial viruses that infect *Prochlorococcus* and *Synechococcus*. Panels A and B represent the noncontracted and contracted tails of myoviruses, respectively. Note that the tails are nonflexible and contain rather conspicuous baseplates and tail fibers. Panels C, D, and E represent siphoviruses that contain long, flexible, noncontractile tails. Note the variability in tail length, tail-terminus structures, and capsid morphology in C and D as compared to E. Panel F shows the icosahedral capsids of podoviruses that contain small, noncontractile tails. All black scale bars are 100 nm. *Photos by M.B. Sullivan, P. Weigele, and B. Ni. Images C and D were originally published in Sullivan et al. (2006)*

Valois, 1993; Proctor, 1997; Wichels et al., 1998; Sullivan et al., 2003) to long filamentous viruses (Middelboe et al., 2003) with particle diameters ranging from 25 nm (*Schizochytrium* single-stranded RNA virus SssRNAV) (Takao et al., 2005) up to ~ 300 nm for a virus that infects a marine phagotrophic protist (Garza and Suttle, 1995). Reported marine viral genome sizes range from 4.4 kilobases (kb) (Tomaru et al., 2004) to 630 kb (Ovreas et al., 2003), with representative genome sequences available from cultured isolates from nearly the extremes of the observed ranges (the 4.4 kb *Heterocapsa circularisquama* virus HcRNAV [Nagasaki et al., 2005a] and the 407 kb Coccolithovirus HeV-86 [Wilson et al., 2005]). Studies targeting genes conserved among members of a viral group (e.g., g20 and g23 of myophages [Fuller et al., 1998; Zhong et al., 2002; Marston and Sallee, 2003; Filee et al., 2005; Short and Suttle, 2005], the RNA polymerase of picorna viruses [Culley et al., 2003], or the DNA polymerase of algal viruses [Chen et al., 1996; Short and Suttle, 2002] and T7-like podophages [Breitbart and Rohwer, 2004]) demonstrate tremendous single-gene diversity even within these restricted groups of viruses. Thus, viral diversity in natural communities is enormous and dynamic as revealed at the levels of morphology, single genes, and whole genome sizes.

Recently, genomic sequencing of marine viral isolates and metagenomic sequencing of marine viral communities has revealed a plethora of previously unknown viruses. Among cultured marine phage genomes, typically between 60% and 80% of the open reading frames show no similarity to any sequences in GenBank (Paul and Sullivan, 2005), while some marine viruses infecting protists have almost no recognizable similarity to extant sequences (Nagasaki et al., 2005b). Furthermore, 65–95% of marine viral metagenomic sequences are not similar to previously described sequences (Breitbart et al., 2002, 2004; Angly et al., 2006; Culley et al., 2006), as opposed to ~ 10% for cellular metagenomic surveys (Tyson et al., 2004; Venter et al., 2004), suggesting that we have only begun to scratch the surface of marine viral sequence diversity.

One of the most striking features of this sequence diversity is an abundance of viral-encoded genes that were previously thought to be restricted to cellular genomes with metabolic capacity. For example, photosynthesis genes, which would seem of little use to something other than a photosynthetic cell, are now thought to be common in cyanophages (Mann et al., 2003; Lindell et al., 2004; Millard et al., 2004; Sullivan et al., 2006). Extensive sequencing efforts on these core photosystem II reaction-center genes show that cyanophages themselves act as genetic reservoirs for their hosts, generating diversity even at

the level of these globally distributed genes (Zeidner et al., 2005; Sullivan et al., 2006). Gene-expression studies on model phage-host pairs show that both messenger RNA (Lindell et al., 2005; Clokie et al., 2006) and protein (Lindell et al., 2005) are produced from viral photosynthesis genes during infection, which suggests that they are functional. Several other so-called "host genes," thought to be remnants of horizontal gene transfer, are present to varying degrees in cyanophages (Chen and Lu, 2002; Mann et al., 2005; Sullivan et al., 2005) and other marine phages (Rohwer et al., 2000; Miller et al., 2003; Lohr et al., 2005). Some of these genes are conserved across multiple phage lineages, such as the photosynthesis and carbon metabolism genes, which suggest that these genes play critical roles during infection, likely augmenting biochemical processes at key metabolic bottlenecks (Figure 2). For this reason, we suggest the term "auxiliary metabolic genes" (AMGs) rather than the potentially misleading term "host genes" when describing these genetic elements.

Traditionally, it was thought that the key role of viruses in microbial food webs was as agents of mortality (up to ~ 50% of prokaryotes are lysed per day by viruses; see reviews in Fuhrman [1999] and Weinbauer [2004]). However, the role of viruses in host metabolism is perhaps even more important. It is now recognized that marine viruses routinely procure AMGs to tap into critical, rate-limiting steps of host metabolism during infection (Sullivan et al., 2006; Angly et al., 2006). Such AMGs are not random evolutionary noise, but rather entrenched parts of viral genomes, akin to nucleotide-metabolism genes long known in coliphages (e.g., ribonucleotide reductase in T4-like phages), and are likely critical to the success of certain viruses in the marine environment. The impact of the role played by viruses is particularly important in environments where viral hosts have global-scale distributions (e.g., the ocean); here, viruses are likely modulating the biogeochemical cycles that run the planet.

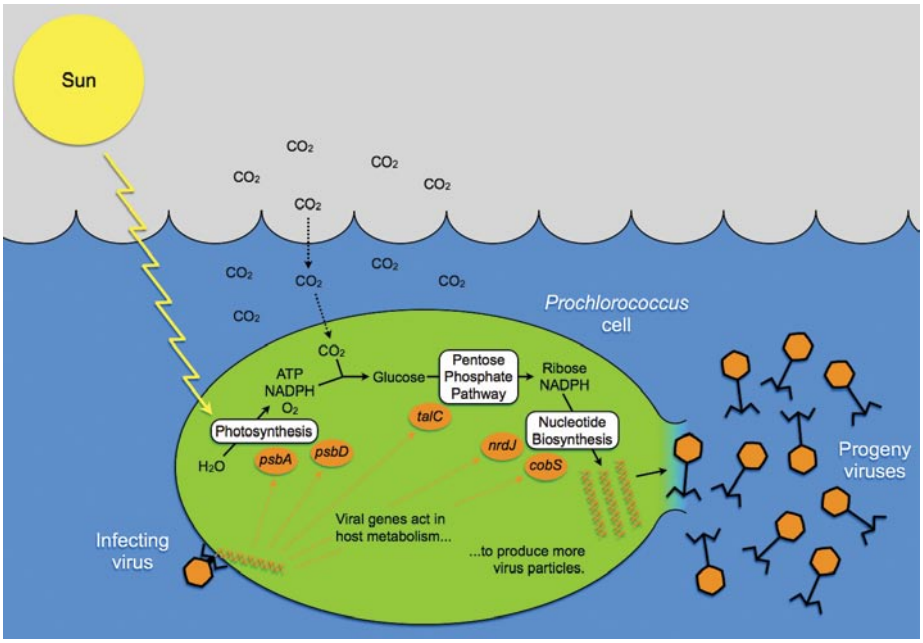


Figure 2. Schematic summarizing the potential roles of cyanophage-encoded "auxiliary metabolic genes" during infection of *Prochlorococcus*, a cyanobacterium. Three cellular metabolic pathways—photosynthesis, the pentose-phosphate pathway, and nucleotide biosynthesis—combine to make nucleotides, critical precursors for DNA replication in both cyanobacteria and their viruses. Infecting viruses often carry genes for photosystem II proteins (*psbA*, *psbD*), transaldolase (*talC*), ribonucleotide reductase (*nrdJ*), and biosynthetic enzymes for making B_{12} (*cobS*), a cofactor of ribonucleotide reductase. When expressed during infection, these genes may augment key steps in cellular metabolism, opening potential bottlenecks to increase nucleotide production, virus genomic DNA replication, and ultimately virus production.

As evidenced by work on photosynthesis genes in cyanophages, the approach of studying model systems in the laboratory is a powerful one. Model systems allow characterization of critical modeling parameters (e.g., extent and mechanisms of host range, burst size, lytic period length), complete genome sequencing to map the capacity to which a given virus might influence ecosystem processes, and, if genetic systems are available, functional assignments for unknown open reading frames. In particular, a synergistic, (meta)genomics-enabled, model-virus-host systems approach can be used to evaluate the ecological roles and the extent of marine-viral diversity. Undoubtedly, as long as the model systems approach relies upon the culturability of organisms (most marine microbes are resistant to culturing), then cautious extrapolation of laboratory results to natural

communities is warranted.

Deep exploration of the diversity and ecosystem function of marine viral communities is a daunting yet exciting task. Tremendous progress has been made using culture-based and signature-gene-based techniques, as well as through metagenomic surveys. Maximizing our interpretation of these rapidly growing metagenomic data sets will require an understanding of cloning and amplification biases of current techniques, and it will also require efforts to isolate and characterize representative viral community members. Future challenges include the development of genetic tools for tracking all major marine groups (e.g., in situ hybridization sequence-based assays using signature genes), the expansion of "snapshot" metagenomic characterizations to evaluate the temporal and spatial dynamics of natural communities, and the development of a robust theoretical framework to enhance our ability to model and predict the impacts of viruses on global ecosystem function.

For further reading on marine viruses, see the following comprehensive reviews: Dunigan et al., 2006; Fuhrman, 1999; Proctor, 1997; Suttle, 2005; Weinbauer, 2004; and Wommack and Colwell, 2000.

ACKNOWLEDGEMENTS

MBS and LRT are partially supported by the Gordon and Betty Moore Foundation and NSF grants to Sallie W. Chisholm.

It is now recognized that marine viruses routinely procure auxiliary metabolic genes to tap into critical, rate-limiting steps of host metabolism during infection.

REFERENCES

Angly, F.E., B. Felts, M. Breitbart, P. Salamon, R.A. Edwards, C. Carlson, A.M. Chan, M. Haynes, S. Kelley, H. Liu, and others. 2006. The marine viriomes of four oceanic regions. *PLoS Biology* 4:e368.

Bergh, Ø., K.Y. Børsheim, G. Bratbak, and M. Heldal, 1989. High abundance of viruses found in aquatic environments. *Nature* 340:467–468.

Breitbart, M., B. Felts, S. Kelley, J.M. Mahaffy, J. Nulton, P. Salamon, and F. Rohwer. 2004. Diversity and population structure of a near-shore marine sediment viral community. *Proceedings of the Royal Society B* 271:565–574.

Breitbart, M., and F. Rohwer. 2004. Global distribution of nearly identical phage-encoded DNA sequences. *FEMS Microbiology Letters* 236: 245–252.

Breitbart, M., P. Salamon, B. Andresen, J.M. Mahaffy, A.M. Segall, D. Mead, F. Azam, and F. Rohwer. 2002. Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences of the United States of America* 99:14,250–14,255.

Chen, F., and J. Lu. 2002. Genomic sequence and evolution of marine cyanophage P60: A new insight on lytic and lysogenic phages. *Applied and Environmental Microbiology* 68:2,589–2,594.

Chen, F., C.A. Suttle, and S.M. Short. 1996. Genetic diversity in marine algal virus communities as revealed by sequence analysis of DNA polymerase genes. *Applied and Environmental Microbiology* 62:2,869–2,874.

Clokie, M.R.J., J. Shan, S. Bailey, Y. Jia, H.M. Krisch, S. West, and N.H. Mann. 2006. Transcription of a 'photosynthetic' T4-type phage during infection of a marine cyanobacterium. *Environmental Microbiology* 8:827–835.

Culley, A.I., A.S. Lang, and C.A. Suttle. 2003. High diversity of unknown picorna-like viruses in the sea. *Nature* 424:1,054–1,057.

Culley, A.I., A.S. Lang, and C.A. Suttle. 2006. Metagenomic analysis of coastal RNA virus communities. *Science* 312:1,795–1,798.

Dunigan, D.D., L.A. Fitzgerald, and J.L. Van Etten. 2006. Phycodnaviruses: A peek at genetic diversity. *Virus Research* 117:119–132.

Filee, J., F. Tetart, C.A. Suttle and H.M. Krisch. 2005. Marine T4-type bacteriophages, a ubiquitous component of the dark matter of the biosphere. *Proceedings of the National Academy of Sciences of the United States of America* 102:12,471–12,476.

Fuhrman, J.A. 1999. Marine viruses: Biogeochemical and ecological effects. *Nature* 399:541–548.

Fuller, N.J., W.H. Wilson, I.R. Joint, and N.H. Mann. 1998. Occurrence of a sequence in marine cyanophages similar to that of T4 g20 and its application to PCR-based detection and quantification techniques. *Applied and Environmental Microbiology* 64:2,051–2,060.

Garza, D.R., and C.A. Suttle. 1995. Large double-stranded DNA viruses which cause the lysis of marine heterotrophic nanoflagellates (*Bodo* sp.) occur in natural marine virus communities. *Aquatic Microbial Ecology* 9:203–210.

Irigoien, X., J. Huisman, and R.P. Harris. 2004. Global biodiversity patterns of marine phytoplankton and zooplankton. *Nature* 429:863–867.

Lindell, D., J.D. Jaffe, Z.I. Johnson, G.M. Church, and S.W. Chisholm. 2005. Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* 438:86–89.

Lindell, D., M.B. Sullivan, Z.I. Johnson, A.C. Tolonen, F. Rohwer, and S.W. Chisholm. 2004. Photosynthesis genes in *Prochlorococcus* cyanophage. *Proceedings of the National Academy of Sciences of the United States of America* 101:11,013–11,018.

Lohr, J.E., F. Chen, and R.T. Hill. 2005. Genomic analysis of bacteriophage fJL001: Insights into its interaction with a sponge-associated alpha-proteobacterium. *Applied and Environmental Microbiology* 71:1,598–1,609.

Mann, N.H., M.R.J. Clokie, A. Millard, A. Cook, W.H. Wilson, P.J. Wheatley, A. Letarov, and H.M. Krisch. 2005. The genome of S-PM2, a "photosynthetic" T4-type bacteriophage that infects marine *Synechococcus* strains. *Journal of Bacteriology* 187:3,188–3,200.

Mann, N.H., A. Cook, A. Millard, S. Bailey, and M.R.J. Clokie. 2003. Marine ecosystems: Bacterial photosynthesis genes in a virus. *Nature* 424:741.

Marston, M.F., and J.L. Sallee. 2003. Genetic diversity and temporal variation in the cyanophage community infecting marine *Synechococcus* species in Rhode Island's coastal waters. *Applied and Environmental Microbiology* 69:4,639–4,647.

Middelboe, M., R.N. Glud, and K. Finster. 2003. Distribution of viruses and bacteria in relation to diagenetic activity in an estuarine sediment. *Limnology and Oceanography* 48:1,447–1,456.

Millard, A., M.R.J. Clokie, D.A. Shub, and N.H. Mann. 2004. Genetic organization of the *psbAD* region in phages infecting marine *Synechococcus* strains. *Proceedings of the National Academy of Sciences of the United States of America* 101:11,007–11,012.

Miller, E.S., J.F. Heidelberg, J.A. Eisen, W.C. Nelson, A.S. Durkin, A. Ciecko, T.V. Feldblyum, O. White, I.T. Paulsen, W.C. Nierman, and others. 2003. Complete genome sequence of the broad-host-range vibriophage KVP40: Comparative genomics of a T4-related bacteriophage. *Journal of Bacteriology* 185:5,220–5,233.

Moebus, K. 1991. Preliminary observations on the concentration of marine bacteriophages in the water around Helgoland. *Helgolander Meeresunters* 45:411–422.

Moebus, K. 1992. Further investigations on the concentration of marine bacteriophages in the water around Helgoland, with reference to the phage-host systems encountered. *Helgolander Meeresunters* 46:275–292.

Nagasaki, K., Y. Shirai, Y. Takao, H. Mizumoto, K. Nishida, and Y. Tomaru. 2005a. Comparison of genome sequences of single-stranded RNA viruses infecting the bivalve-killing dinoflagellate *Heterocapsa circularisquama*. *Applied and Environmental Microbiology* 71:8,888–8,894.

Nagasaki, K., Y. Tomaru, Y. Takao, K. Nishida, Y. Shirai, H. Suzuki, and T. Nagumo. 2005b. Previously unknown virus infects marine diatom. *Applied and Environmental Microbiology* 71:3,528–3,535.

Ovreas, L., D. Bournde, R.A. Sandaa, E. Casamayor, S. Benlloch, V. Goddard, G. Smerdon, M. Heldal, and T.F. Thingstad. 2003. Response of bacterial and viral communities to nutrient manipulations in seawater mesocosms. *Aquatic Microbial Ecology* 31:109–121.

Paul, J.H., and M.B. Sullivan. 2005. Marine phage genomics: What have we learned? *Current Opinion in Biotechnology* 16:299–307.

Proctor, L.M. 1997. Advances in the study of marine viruses. *Microscopy Research and Technique* 37:136–161.

Rohwer, F., A.M. Segall, G. Steward, V. Seguritan, M. Breitbart, F. Wolven, and F. Azam. 2000. The complete genomic sequence of the marine phage Roseophage SIO1 shares homology with non-marine phages. *Limnology and Oceanography* 42:408–418.

Short, C.M., and C.A. Suttle. 2005. Nearly identical bacteriophage structural gene sequences are widely distributed in marine and freshwater environments. *Applied and Environmental Microbiology* 71:480–486.

Short, S.M., and C.A. Suttle. 2002. Sequence analysis of marine virus communities reveals that groups of related algal viruses are widely distributed in nature. *Applied and Environmental Microbiology* 68:1,290–1,296.

Spencer, R. 1955. A marine bacteriophage. *Nature* 175:690.

Sullivan, M.B., M.L. Coleman, P. Weigele, F. Rohwer, and S.W. Chisholm. 2005. Three *Prochlorococcus* cyanophage genomes: Signature features and ecological interpretations. *PLoS Biology* 3:790–806.

Sullivan, M.B., D. Lindell, J.A. Lee, L.R. Thompson, J.P. Bielawski, and S.W. Chisholm. 2006. Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biology* 4:e234.

Sullivan, M.B., J.B. Waterbury, and S.W. Chisholm. 2003. Cyanophages infecting the oceanic cyanobacterium *Prochlorococcus*. *Nature* 424:1,047–1,051.

Suttle, C.A. 2005. Viruses in the sea. *Nature* 437:356–361.

Takao, Y., K. Nagasaki, K. Mise, T. Okuno, and D. Honda. 2005. Isolation and characterization of a novel single-stranded RNA virus infectious to a marine fungoid protist, *Schizochytrium sp.* (Thraustochytriaceae, Labyrinthulea). *Applied and Environmental Microbiology* 71:4,516–4,522.

Thompson, J.R., S. Pacocha, C. Pharino, V. Klepac-Ceraj, D.E. Hunt, J. Benoit, R. Sarma-Rupavtarm, D.L. Distel, and M.F. Polz. 2005. Genotypic diversity within a natural coastal bacterioplankton community. *Science* 307:1,311–1,313.

Tomaru, Y., N. Katanozaka, K. Nishida, Y. Shirai, K. Tarutani, M. Yamaguchi, and K. Nagasaki. 2004. Isolation and characterization of two distinct types of HcRNAV, a single-stranded RNA virus infecting the bivalve-killing microalga *Heterocapsa circularisquama*. *Aquatic Microbial Ecology* 34:207–218.

Tyson, G.W., J. Chapman, P. Hugenholtz, E.E. Allen, R.J. Bram, P.M. Richardson, V.V. Solovyev, E.M. Rubin, D.S. Rokhsar, and J.F. Banfield. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428:37–43.

Venter, J.C., K. Remington, J.F. Heidelberg, A.L. Halpern, D. Rusch, J.A. Eisen, D. Wu, I. Paulsen, K.E. Nelson, W. Nelson, and others. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304:66–74.

Waterbury, J.B., and F.W. Valois. 1993. Resistance to co-occurring phages enables marine *Synechococcus* communities to coexist with cyanophages abundant in seawater. *Applied and Environmental Microbiology* 59:3,393–3,399.

Weinbauer, M.G. 2004. Ecology of prokaryotic viruses. *FEMS Microbiology Reviews* 28:127–181.

Wichels, A., S.S. Biel, H.R. Gelderblom, T. Brinkhoff, G. Muyzer, and C. Schuett. 1998. Bacteriophage diversity in the North Sea. *Applied and Environmental Microbiology* 64:4,128–4,133.

Wilson, W., D.C. Schroeder, M.J. Allen, M.T.G. Holden, J. Parkhill, B.G. Barrell, C. Churcher, N. Hamlin, K. Mungall, H. Norbertczak, and others. 2005. Complete genome sequence and lytic phase transcription profile of a coccolithovirus. *Science* 5737:1,090–1,092.

Wilson, W.H., I.R. Joint, N.G. Carr, and N.H. Mann. 1993. Isolation and molecular characterization of five marine cyanophages propagated on *Synechococcus* sp. strain WH7803. *Applied and Environmental Microbiology* 59:3,736–3,743.

Witman, J.D., R.J. Etter, and F. Smith. 2004. The relationship between regional and local species diversity in marine benthic communities: A global perspective. *Proceedings of the National Academy of Sciences of the United States of America* 101:15,664–15,669.

Wommack, K.E., and R.R. Colwell. 2000. Virioplankton: Viruses in aquatic ecosystems. *Microbiology and Molecular Biology Reviews* 64:69–114.

Worden, A.Z. 2006. Picoeukaryote diversity in coastal waters of the Pacific Ocean. *Aquatic Microbial Ecology* 43:165–175.

Zeidner, G., J.P. Bielawski, M. Shmoish, D.J. Scanlan, G. Sabehi, and O. Beja. 2005. Potential photosynthesis gene recombination between *Prochlorococcus* and *Synechococcus* via viral intermediates. *Environmental Microbiology* 7:1,505–1,513.

Zhong, Y., F. Chen, S.W. Wilhelm, L. Poorvin, and R.E. Hodson. 2002. Phylogenetic diversity of marine cyanophage isolates and natural virus communities as revealed by sequences of viral capsid assembly protein gene g20. *Applied and Environmental Microbiology* 68:1,576–1,584.

APPENDIX F

Choreography of the transcriptome, photophysiology, and cell cycle of a minimal photoautotroph, *Prochlorococcus* (Zinser et al., *PLoS One*, 2009)

OPEN ACCESS Freely available online

PLoS one

Choreography of the Transcriptome, Photophysiology, and Cell Cycle of a Minimal Photoautotroph, *Prochlorococcus*

Erik R. Zinser[1,2], Debbie Lindell[1,3], Zackary I. Johnson[1,4], Matthias E. Futschik[5,6], Claudia Steglich[1,7], Maureen L. Coleman[1], Matthew A. Wright[8], Trent Rector[8], Robert Steen[8], Nathan McNulty[1], Luke R. Thompson[9], Sallie W. Chisholm[1,9]*

1 Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, **2** Department of Microbiology, University of Tennessee, Knoxville, Tennessee, United States of America, **3** Faculty of Biology, Technion-Israel Institute of Technology, Haifa, Israel, **4** Department of Oceanography, University of Hawaii, Honolulu, Hawaii, United States of America, **5** Institute of Theoretical Biology, Humboldt University, Berlin, Germany, **6** Center for Molecular and Structural Biomedicine, University of Algarve, Faro, Portugal, **7** Institute of Biology III, University of Freiburg, Freiburg, Germany, **8** Department of Genetics, Harvard Medical School, Boston, Massachusetts, United States of America, **9** Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America

Abstract

The marine cyanobacterium *Prochlorococcus* MED4 has the smallest genome and cell size of all known photosynthetic organisms. Like all phototrophs at temperate latitudes, it experiences predictable daily variation in available light energy which leads to temporal regulation and partitioning of key cellular processes. To better understand the tempo and choreography of this minimal phototroph, we studied the entire transcriptome of the cell over a simulated daily light-dark cycle, and placed it in the context of diagnostic physiological and cell cycle parameters. All cells in the culture progressed through their cell cycles in synchrony, thus ensuring that our measurements reflected the behavior of individual cells. Ninety percent of the annotated genes were expressed, and 80% had cyclic expression over the diel cycle. For most genes, expression peaked near sunrise or sunset, although more subtle phasing of gene expression was also evident. Periodicities of the transcripts of genes involved in physiological processes such as in cell cycle progression, photosynthesis, and phosphorus metabolism tracked the timing of these activities relative to the light-dark cycle. Furthermore, the transitions between photosynthesis during the day and catabolic consumption of energy reserves at night— metabolic processes that share some of the same enzymes — appear to be tightly choreographed at the level of RNA expression. In-depth investigation of these patterns identified potential regulatory proteins involved in balancing these opposing pathways. Finally, while this analysis has not helped resolve how a cell with so little regulatory capacity, and a 'deficient' circadian mechanism, aligns its cell cycle and metabolism so tightly to a light-dark cycle, it does provide us with a valuable framework upon which to build when the *Prochlorococcus* proteome and metabolome become available.

Citation: Zinser ER, Lindell D, Johnson ZI, Futschik ME, Steglich C, et al. (2009) Choreography of the Transcriptome, Photophysiology, and Cell Cycle of a Minimal Photoautotroph, *Prochlorococcus*. PLoS ONE 4(4): e5135. doi:10.1371/journal.pone.0005135

Editor: Francisco Rodriguez-Valera, Universidad Miguel Hernandez, Spain

Received December 23, 2008; **Accepted** January 19, 2009; **Published** April 8, 2009

Funding: This work was supported by grants from the NSF, DOE, the Seaver Foundation, and the Gordon and Betty Moore Foundation to S.W.C, from the NSF and the University of Tennessee to E.R.Z., from the NSF, NOAA, and the University of Hawaii to Z.I.J., and from the DFG (SPP 1258) to C.S. DL is a Shillman fellow. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: chisholm@mit.edu

Introduction

The unicellular cyanobacterium *Prochlorococcus* is believed to be the most abundant photosynthetic organism on Earth [1]. It is also the smallest oxygenic phototroph, both in physical size (0.6 microns in diameter) and genome size. The latter ranges from 1.64–2.68 Mbp in a set of strains that span the currently known phylogenetic diversity of this group [2]. The streamlined genome appears to be accompanied by a reduction in regulatory capacity. Strain MED4, for example, contains only five sigma factors, five sensor histidine kinases, and seven response regulators, considerably fewer than that found in other bacteria [3]. The relative number of non-coding RNAs is comparable to that found in other bacteria, however [4], which suggests an unusual regulation portfolio in this organism. Rapid shifts in temperature, salinity, pH, and other physical variables are rare in the ocean environment, and nutrients are typically maintained at extremely low concentrations, except during deep mixing events in seasonal environments. The overall reduction in regulatory capacity could be viewed as streamlining for life in a relatively static environment.

Life in the nutrient-poor open ocean is not devoid of dynamism, however. Sunlight, the energy source for *Prochlorococcus*, undergoes a regular and dramatic variation in supply each day. It is not surprising, therefore, to find that cellular metabolism has been shaped by this diel energy flux. Carbon fixation in *Prochlorococcus* has been shown to occur exclusively during the day, with approximately 2/3 of the total carbon accumulation occurring before mid-day [5,6]. Other photosynthetic parameters, such as photochemical efficiency of photosystem II (F_v/F_m), quantum yield of chlorophyll fluorescence, maximum quantum yield of carbon

fixation, and concentration of the carotenoid accessory pigment zeaxanthin, also showed strong diel variation in prior studies [5,6].

The expression of a number of photosynthesis genes are known to display periodicity over a diel light/dark cycle in *Prochlorococcus*. Transcripts of genes encoding photosystem II's D1 (*psbA*), D2 (*psbD*), and CP43 (*psbC*), for example, peak in abundance at subjective mid-day, while the major light-harvesting complex (*pcbA*, or *pcb* in strain MED4) has two maxima, one at sunrise, and one at sunset [7]. Expression of the *rbcL* gene encoding the large subunit of the Rubisco, parallels strongly with the carbon fixation rate and maximum quantum yield of carbon fixation, exhibiting a pronounced maximum at sunrise and a dramatic decrease in the afternoon [5,8].

The cell cycles of *Prochlorococcus* cells cultured on light dark cycles are tightly synchronized [9–12]. In populations with mean generation times of one day or longer, which is typical under most conditions [11,13], DNA synthesis occurs during the afternoon, and cell division — in those cells that divide — occurs only in the late afternoon or early evening [11,14,15]. In cases where populations double more than once per day, the second round of division takes place within hours of the first [13]. Not surprisingly, expression of genes involved in initiating cell division (*ftsZ*) and DNA replication (*dnaA*) varies significantly over the light/dark cycle in synchronized cultures, and are maximal during the S phase [16].

Given the tight cell cycle synchrony on light/dark cycles, and periodicity of so many other cellular functions in *Prochlorococcus*, one might suspect that these processes are regulated through coupling to a circadian oscillator, as is typical of other cyanobacteria. For example, transcription of much of the genome in freshwater cyanobacteria, and the regulation of key physiological processes in freshwater and marine cyanobacteria, have been found to be under the control of a circadian clock [17–21]. Three components, KaiA, KaiB, and KaiC, are necessary and sufficient for the clock to function [21,22], and transmission of the clock signal to the genome is believed to occur through the SasA-RpaA two-component regulatory system [23] or SasA-independent changes in DNA topology [21,24]. Light-dependent entrainment of the clock appears to work through CikA, which modifies the phosphorylation state of KaiC [25].

While *Prochlorococcus* contains the clock genes *kaiB* and *kaiC* [12], and they have periodic expression on a light/dark cycle [12], it lacks *kaiA*. The latter is believed to be an essential component of the cyanobacterial clock as it is involved in phosphorylating KaiC, and in helping the clock keep time in absence of light-dark cues. Importantly, whereas cyanobacteria that contain *kaiA* maintain periodic expression under constant light conditions, *Prochlorococcus* does not [12]. Furthermore, several key regulators of the clock that are involved in light-dark entrainment (e.g. CikA) are missing in *Prochlorococcus* [12], suggesting either that *Prochlorococcus* does not have a clock, or that it functions in a different way.

The extremely tight synchrony of cell division in *Prochlorococcus* when grown on a light/dark cycle, its streamlined genome, and its apparent limitations *vis a vis* a functioning circadian oscillator, motivated us to undertake an in-depth analysis of the coordination of the transcriptome, cell cycle, and photophysiology in this cell. The questions driving our study were as follows: What fraction of the entire genome is expressed under optimal growth conditions on a light-dark cycle, and what fraction of those expressed genes are periodic? What is the temporal relationship between the timing of transcription of key genes, and the physiological processes they are associated with? What genes are transcribed at similar times in the cycle, and does this clustering tell us anything about metabolic partitioning? Finally, what can we learn about the global regulation of diel periodicity in gene expression, particularly as this cell seems to lack a circadian clock?

Results and Discussion

Prochlorococcus strain MED4, a member of the high-light adapted clade of *Prochlorococcus* that dominates surface waters over much of the mid-latitude oceans [26] was used for this study. It has one of the smallest genomes of all cultured *Prochlorococcus* strains, synchronizes tightly to a light dark cycle, and can achieve a growth rate of one doubling per day under optimal conditions. The doubling times of the replicate cultures used in this study were 1.1 and 1.0 days, and thus the cells within the population progressed through the cell cycle in synchrony. The important consequence is that our population-level measurements of gene expression and cell physiology approximate what is happening in an individual cell. As a result, the periodicity in the global transcriptome was very well defined and reproducible over both days of sampling in both of the replicate cultures (Figure 1, and see Table S1 for expression data), forming a solid database for all of our analyses.

General features of the transcriptome and its response to the light dark cycle

Overall, 89% of the total 1698 analyzed protein-coding genes in this cell were expressed at detectable levels (see Materials and Methods section) over the photocycle. The remaining genes include 27 that have been shown to be upregulated in response to nutrient and light stress, as well as phage infection [27–30] —just a few of the stressors that *Prochlorococcus* cells are likely to experience in the oceans. We hypothesize that the remaining genes with undetectable expression may play similar roles. All *Prochlorococcus* strains sequenced to date share 1273 gene clusters, constituting a well-defined set of 'core' genes for this group, which is also supported by analyses of metagenomic databases [2,31–33]. *Prochlorococcus* MED4 contains an additional 615 so-called 'flexible' gene clusters, which are found in some, but not all strains of *Prochlorococcus*. Flexible genes are often located in hypervariable genomic islands thought to play a role in adaptation to specific environments. Since the core genes encode basic metabolic

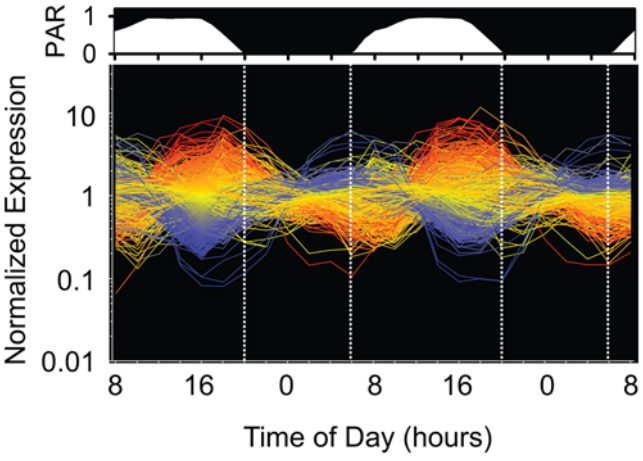


Figure 1. Relative RMA-normalized expression levels of all annotated open reading frames in MED4 over a two-day diel. Relative PAR (photosynthetically available radiation) over the experiment is represented above the expression patterns. Each line represents one of the 1698 unique open reading frames (line colors are arbitrary). Vertical dotted lines denote light/dark transitions.
doi:10.1371/journal.pone.0005135.g001

processes [2,34] whereas the 'flexible' genes are more specialized, one might expect that the 'core' genes would be disproportionately expressed relative to the flexible genes under the optimal growth conditions of our experiments. We found a marginal difference: 91% of the 1288 core genes compared with 83% of the 410 flexible genes were expressed.

Qualitative inspection shows that most of the genes display periodic expression, with a single maximum and minimum per 24 hour photoperiod (Figure 1). Fourier analysis revealed that 91% (with a false discovery rate (FDR) of less than 0.1) of the expressed protein-encoding genes exhibited significant periodicity. In contrast to the protein coding genes, only 68% of non-coding RNAs (excluding tRNA and ribosomal RNA genes) and 67% of antisense RNAs were periodic. Many of the aperiodic ncRNAs are "house keeping" genes such as *mpB*, *ffs* and *ssrA* (Table S1). Probes derived from the intergenic regions displayed a considerably lower percentage of periodic expression (31%). The intergenic probe sets that exhibited periodicity may correspond to 5′/3′ untranslated regions, genes missed in the initial genome annotation, or short functional RNAs [4]. Of the "flexible" genes, 90% of those expressed were periodic, including those in genomic islands. Thus at the transcriptional level, the flexible genome, and even genomic islands, have similar characteristics as the core genome, lending support to the hypothesis that the flexible genome and genomic islands are physiologically important.

We next looked at the overall features and timing of the expression patterns of the periodically expressed genes. For most genes, peak expression was at the onset of either subjective sunrise or sunset (Figure 1). Quantitative analyses (see Methods) confirmed that the distribution of the time of maximum RNA abundance over the photoperiod for all of the periodic genes was largely bimodal, with most genes peaking in expression within a few hours of subjective sunrise (06:00) or sunset (20:00) (Figure S1). Despite this clustering around dawn and dusk, every hour in the 24 hour photoperiod was the time of peak expression of at least a few genes (Figure S1). To identify the predominant patterns of diel periodicity, we performed "soft clustering" analysis (see Materials and Methods) of the transcriptome. Sixteen clusters of genes could be identified as having similar transcriptome periodicities (Table 1, Figure S2). The size of the clusters ranged from 22 to 138 (average 88) genes and peak transcription levels of the clusters were spread fairly evenly over the photocycle, with the exception of clusters 12 and 13, and 14 and 15, which had peak expression times less than one half hour apart. The gene content of these clusters and their relationships form the heart of the analysis of transcriptome coordination presented below.

Cell growth and the cell division cycle

The tight synchrony of the cells in the cultures was reflected in a number of the measured variables. The growth of individual cells

Table 1. Characteristics of the gene clusters found to be periodic (1–16), aperiodic (17), and non-expressed (18), showing the time of their peak expression (note that h = 0, is 4 hours after the onset of dark in a 14:10 light-dark cycle), and the subcategory of genes enriched in each cluster.

Cluster	# Genes	Mean Fourier Score	Mean peak time (h)	Cyanobase Subcategories enriched	Subcategory genes: Enriched / Total	Enrichment FDR
1	57	15.69	8.3±0.7	8.5 Photosystem I	9/22	1.50E-09
				8.6 Photosystem II	8/22	2.50E-05
2	52	14.7	9.6±0.8			
3	23	15.02	12.5±0.8	8.3 Cytochrome b6/f	3/7	0.0059
				8.6 Photosystem II	8/22	3.70E-08
4	62	14.97	15.8±0.7	8.3 Cytochrome b6/f	3/7	0.073
5	120	15.96	17.5±0.4	8.9 Respiratory terminal oxidases	3/3	0.019
				13.3 Degradation of proteins, peptides, and glycopeptides	5/15	0.071
6	138	15.06	18.6±0.5	9.2 Purine ribonucleotide biosynthesis	7/18	0.0049
7	121	15.54	20.1±0.4	5.2 Nitrogen metabolism	4/8	0.087
8	90	13.97	21.0±0.6	4.3 Chaperones	7/14	0.00023
9	99	13.94	22.4±0.5	12.2 RNA synthesis, modification, and DNA transcription	6/23	0.035
10	77	13.47	0.3±0.6			
11	91	13.71	1.6±0.4			
12	111	13.96	3.0±0.5			
13	110	15.23	3.4±0.4	13.2 Ribosomal proteins	45/53	3.70E-41
14	22	12.95	4.6±0.8			
15	107	14.76	4.7±0.5			
16	125	15.97	5.5±0.5	8.1 ATP synthase	8/8	8.40E-08
				8.2 CO2 metabolism	7/9	1.80E-05
17	173	7.76	N/A			
18	180	8.08	N/A	2.6 Menaquinone and ubiquinone	6/9	0.00045

The false discovery rate (FDR) of the enrichment is also shown.

doi:10.1371/journal.pone.0005135.t001

(as measured by forward light scatter, a proxy for size) began at dawn, and ended two hours before dark, when the cells began to divide (Figure 2A). Cell number increased in the cultures over the dark period, such that all of the cells had divided by sunrise, i.e. the culture had doubled. DNA synthesis (S phase) began approximately six hours after dawn and was complete by the middle of the night (Figure 2B). The G1 and G2 phases of the cell cycle lead and followed the S phase, with some overlap, but on the whole the population displayed remarkable and reproducible synchrony, both between the replicate cultures and over sequential 24 hour periods.

The temporal specificity of DNA synthesis and cell division during the photocycle was matched by the expression of the genes responsible for these activities. MED4 lacks orthologs to most of the 15 protein components of the cell division machinery ("divisome") of *E. coli* [35], but those it does have were in general maximally expressed prior to the onset of septation (Figure 2C, Table S2). Transcript levels of *ftsZ*, for example, which encodes the cytoplasmic septal Z ring, peaked 4 hours before sunset at the time of the S-phase maximum (Figure 2C), consistent with prior studies [16]. A trio of proteins, MinC, MinD, and MinE, function to establish the location of FtsZ ring formation [36], thus it is not surprising that transcript abundance for *minD* (Table S2) and *minE* (Figure 2C and Table S2) exhibited strong periodicity with a pattern similar to *ftsZ* in our experiment (Figure 2C). The pattern for *minC* was also periodic, though not as strong (Table S2). Expression of *ftsI* and *ftsW*, which together synthesize the septal peptidoglycan once recruited to the Z-ring [35], peaked 1–2 hours after *ftsZ* (Figure 2C and Table S2), timing that is consistent with that of *E. coli* [35]. In contrast, the two paralogs of *ftsI* and *ftsW*, *pbp2* and *rodA* respectively, were expressed aperiodically (Table S2), which is consistent with their function in the synthesis of the cell wall during cell growth rather than division [35]. Two other predicted members of the cell division apparatus, *mraW* and *amiC* - encoding an S-adenosyl-methionine-dependent methyltransferase and a periplasmic amidase, respectively [35] - had undetectable expression or peak expression at 03:00, respectively (Table S2), leaving their role unclear.

As DNA synthesis occurred at a discrete period in the light/dark cycle, so did the peak abundance of the genes involved in this process. Initiation of chromosomal replication involves proteins DnaA and DnaB (helicase), thus it is not surprising that their transcripts accumulated 2–6 hours prior to the onset of DNA replication, and were maximally-abundant at the peak of S phase (Figure 2D and Table S3), confirming prior isolated studies of *dnaA* expression in *Prochlorococcus* [16]. Genes involved in initiation as well as elongation phases of DNA polymerization were likewise maximally abundant during the S phase. This includes 4 out of the 5 genes encoding DNA polymerase III (e.g. *dnaE*, Figure 2D), as well as those that encode gyrase (*gyrA*, *gyrB*), primase (*dnaG*), ligase (*ligA*), and the single-stranded binding protein (*ssb*) (Table S3). *polA*, encoding DNA polymerase I, was the key exception, as it showed weak diel periodicity with a night-time maximum (Figure 2D). The weak periodicity may reflect *polA*'s additional role in DNA repair [37], as DNA photodamage during the daytime is likely to be a significant challenge to this high-light adapted strain.

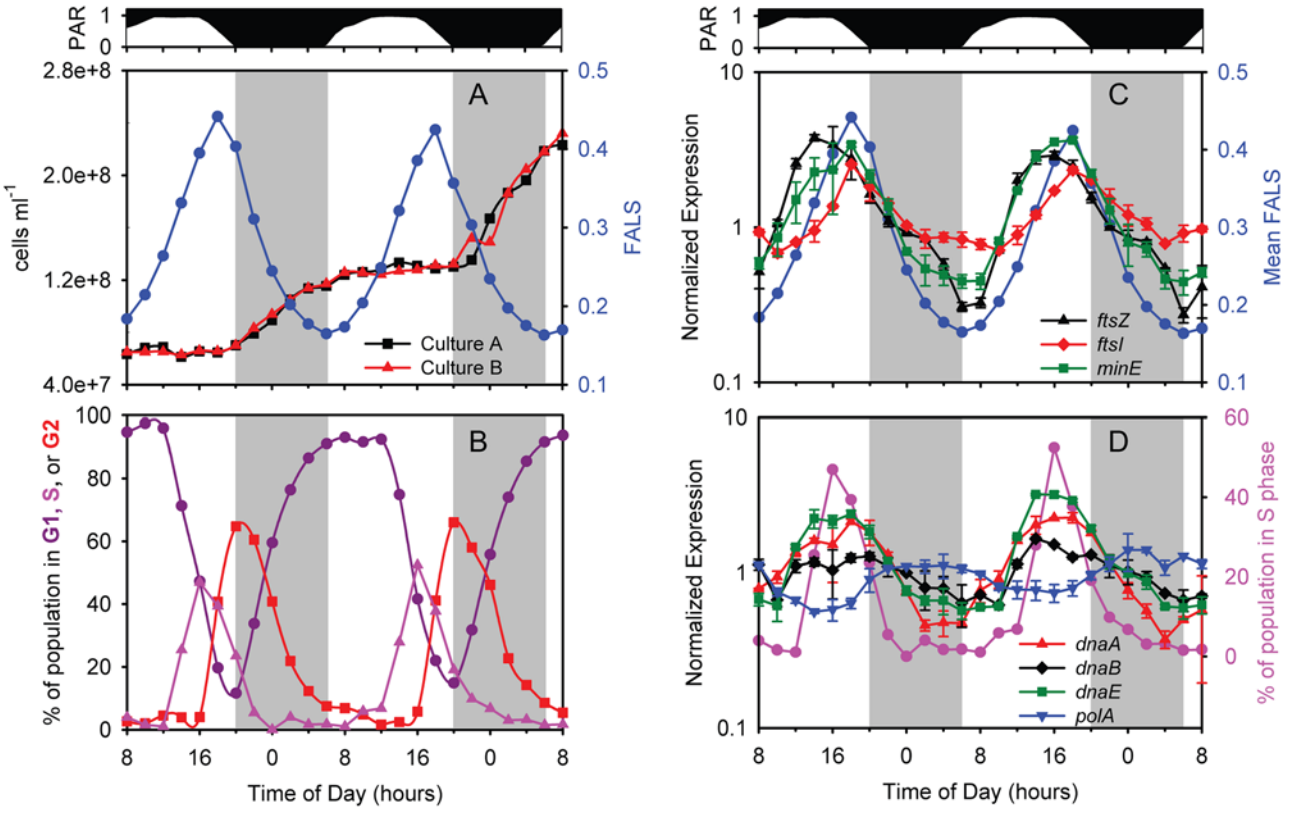


Figure 2. Cell cycle properties of the MED4 cultures during the experiment. (A) Cell abundance for the two individual cultures and mean forward angle light scattering (FALS), a surrogate for cell size, and (B) percentage of cells in G1, S (DNA synthesis), or G2 phase at each time point are shown. Expression time course of (C) cell division genes *ftsZ*, *ftsI*, *minE* and (D) DNA replication genes *dnaA*, *dnaB*, *dnaE*, and *polA*. Mean FALS (C, blue circles) or the percent of the population in S phase (D, pink circles) is shown for comparison. Error bars represent one standard deviation of the mean for the replicate cultures.
doi:10.1371/journal.pone.0005135.g002

This direct comparison of the timing of cell division and DNA synthesis with the transcriptome reveals a rather striking choreography of cell cycle progression in *Prochlorococcus*. With few exceptions, the expression of cell cycle-related genes is periodic in a way that suggests a "just-in-time" transcription of genes encoding key steps in the cells progression through the cycle. We do not know if this results in a "just-in-time" translation of the mRNAs into protein, and if so whether such a boost in protein abundance could play a role in triggering these cell cycle events. None the less, the close match between the periodicity of the genes responsible for cell cycle progression and progression itself is striking.

Photosynthesis

As one would expect, cell-normalized photosynthetic rate (P^{cell},) directly followed the diel light cycle with peak rates of 8.8 ± 0.5 fg C cell^{-1} hr^{-1} occurring at mid-day (Figure 3A). Integrated photosynthesis over the 24 hour period averaged 82.5 ± 0.5 fg C cell^{-1} d^{-1} for the two days, which represents the daily gross photosynthesis per cell. *Prochlorococcus* has an average cellular carbon content of ~53 fg cell^{-1} [38], thus this would be the net carbon fixation needed in a day for a cell to double. Since these cultures are doubling once per day, one can conclude from this that the cell respires and/or excretes roughly a third of the carbon it fixes through photosynthesis.

Also as expected, photophysiological parameters were not static over the light dark cycle. For example, both P^{cell}_{max} (maximum light-saturated photosynthesis – a measure of photosynthetic capacity) and α^{cell}_{max} (maximal instantaneous light utilization – a measure of photosynthetic efficiency, see Materials and Methods) had strong periodicities (Figure 3B), with the former reaching a maximum at mid-day, and the latter reaching one closer to dusk. Minima for the two occurred right before dawn. Because P^{cell}_{max} and α^{cell}_{max} were not in exact phase and did not have the same changes in amplitude, the light saturation index [39], E_k ($P^{cell}_{max}/\alpha^{cell}_{max}$, which is a measure of the maximum light intensity that can be used by the cells) also oscillated with the diel cycle (Figure 3C). It is particularly noteworthy that E_k was highest when photons were most abundant (Figure 3C), indicating that the photosynthetic machinery of the cell is running near its maximal capacity (i.e. $P^{cell}/P^{cell}_{max}\approx1$) for a large portion of the day (Figure 3D), even though optimal light utilization efficiency

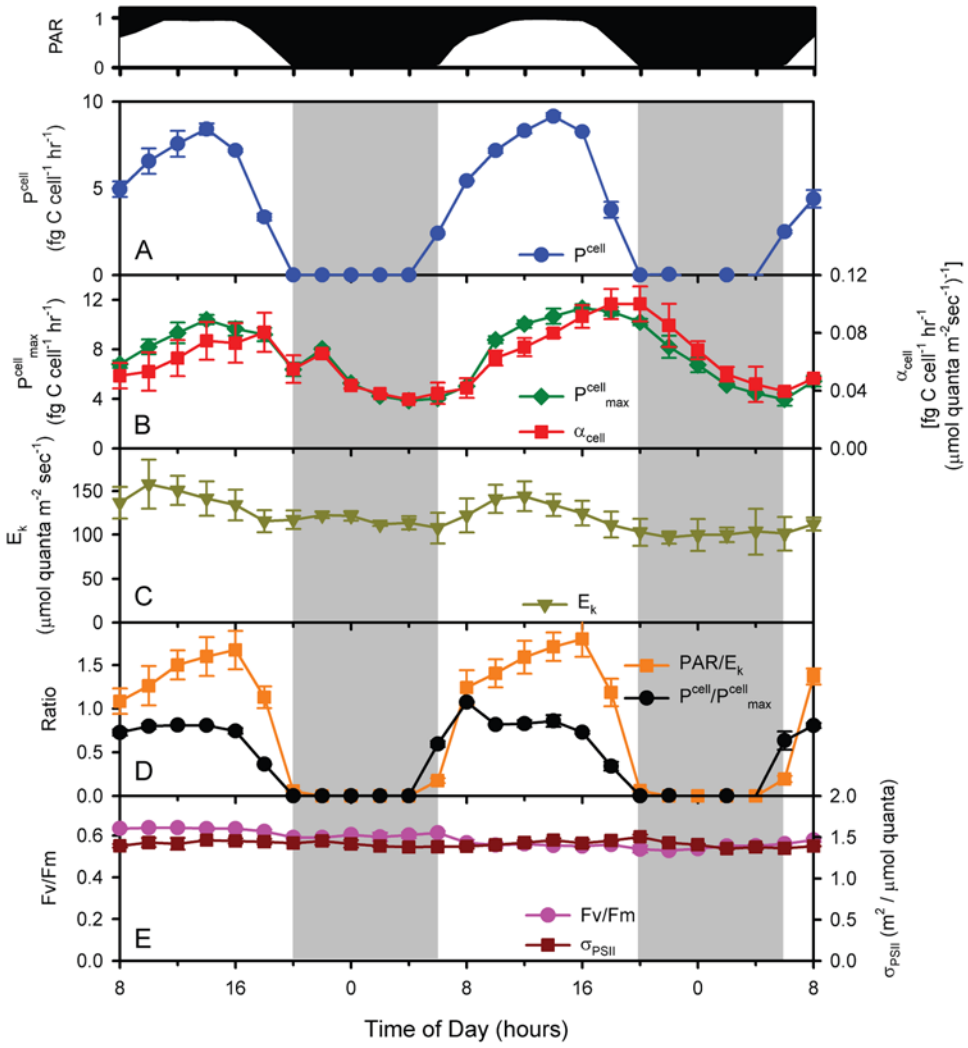


Figure 3. Photosynthesis parameters measured over the time course of the experiment. (A) P^{cell} – cell-normalized photosynthetic rate; (B) P^{cell}_{max} – maximum light-saturated photosynthesis, α^{cell}_{max} – maximal instantaneous light utilization; (C) E_k – maximum light intensity that can be used by the cells; (D) P^{cell}/P^{cell}_{max}, PAR/E_k; and (E) Fv/Fm and σ_{PSII} are shown (see text for details).
doi:10.1371/journal.pone.0005135.g003

(α^{cell}_{max}) may not be achieved. Achieving this maximal energy throughput throughout the day comes at the cost of not using all available photons (i.e. PAR>E_k) for most of the day (Figure 3D), even though this excess light energy does not cause photodamage as evidenced by the invariant Fv/Fm and σ_{PSII} (Figure 3E). Overall, this may be an effective strategy to minimize excess photosynthetic capacity, and the respiratory costs associated with it, thus realizing the highest overall photosynthesis/respiration ratio even though there is additional energy available that could be used. In addition, other sinks for photosynthetic reducing power beyond carbon reduction likely represent important pathways [40]. Thus the diel variability in these photosynthesis parameters demonstrates that although light availability is the proximal factor regulating photosynthetic rates, the photophysiology of MED4 is continually acclimating over the diel cycle and/or cell cycle to maintain balance between light availability and efficiency of utilization.

Given this finely tuned physiology, it is not surprising that the expression of many of the underlying genes had strong periodicity in other cyanobacteria [41–44] as well as in MED4 (this study). Periodicity patterns of photosynthesis genes fell into 4 clusters (Table 1). Expression of approximately half of photosystem (PS) II genes, including reaction center genes *psbA* and *psbD* (encoding D1 and D2 respectively), as well as *psbC* (CP43) and *psbF*, co-varied with light intensity, with maxima at mid-day, and minima in the middle of the night (Figure 4A, Table S4), consistent with patterns observed by Garczarek et al (2001) and Holtzendorff et al. (2008) in their diel study of selected genes in *Prochlorococcus* PCC 9511. F_v/F_m, a measure of the efficiency of PSII, did not change over the course of the experiment (Figure 3E) indicating that the differential expression of photosystem II genes and subsequent protein turnover and reaction center repair was able to mitigate against the damage to PSII [45–47]. This is further supported by only minor (<10%) diel changes in the PSII cross-section (σ_{PSII}) (Figure 3E). Together, these observations suggest that MED4 may maintain PSII reaction center integrity through changes in gene expression.

A second group of PSII genes including components of the reaction center (*psbK*, *psbO* and *psbH*) peaked earlier in the day —

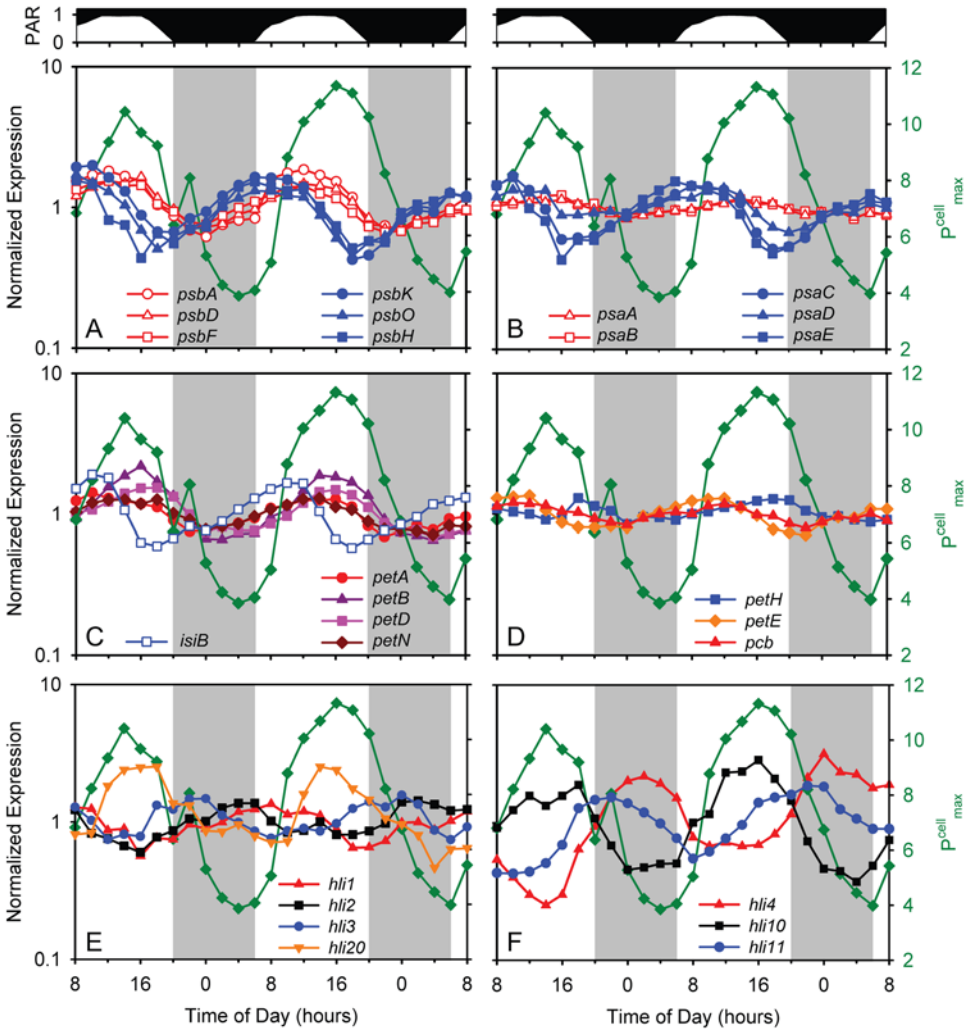


Figure 4. Expression time course of photosynthesis genes. (A) Photosystem II genes *psbA*, *psbD*, *psbF*, *psbH*, *psbK*, and *psbO*; (B) photosystem I genes *psaA*, *psaB*, *psaD*, *psaE*, and *psaC*; (C) Photosynthetic electron transport chain genes *petA*, *petN*, *petB*, *petD*, and *isiB*; and (D) low-periodicity photosynthesis genes *pcb*, *petE*, and *fnr*; and representative high light inducible protein (HLIP)- encoding genes of the (E) single copy - *hli1*, *hli2*, *hli3*, and *hli20* - and (F) multiple copy – *hli4*, *hli10*, and *hli11* - class, with peak abundances at different times over the photocycle are shown. For comparison, P^{cell}_{max} (see Figure 3B) is also reported (green). For clarity, error bars representing sample-to-sample variability in gene expression are not shown.

doi:10.1371/journal.pone.0005135.g004

mid-morning — during the G1 phase (Figure 4A, Table S4), and is likely tied to the *de novo* synthesis of reaction centers after cell division [48]. PSI genes largely peak in expression at the same time, except for *psaA* and *psaB* (both PSI core proteins) which display a very low amplitude of expression, with a peak in mid-afternoon (Figure 4B). These results might lead one to hypothesize that the reaction center core proteins of both PSII and PSI, as well as about half of the proteins associated with PSII, are responding directly to light intensity [49,50] while the remaining PSII and PSI genes are more closely tied to cell cycle processes (i.e. biomass production beginning at sunrise).

Other genes encoding proteins involved in photosynthesis also displayed periodic expression. For example, genes associated with the photosynthetic electron transport chain (PETC) including *isiB* (encoding flavodoxin), and *petA*, *petB*, *petD*, and *petN* (encoding subunits of the cytochrome b_6f complex) have maxima just prior to or during the period of maximum light intensity (Figure 4C). This suggests that either the components of the PETC are becoming damaged because of oxidative stresses, such as with PSII, or that MED4 is up-regulating the throughput capacity of PETC in response to elevated excitation pressure. The timing of the maximum in maximum photosynthetic capacity (P^{cell}_{max}) is coincident with the expression maximum of many PETC genes suggesting that PETC throughput increases shortly after noon. It has been shown in the field and laboratory, via changes in the turnover time of PSII ($1/\tau_{PSII}$), that phytoplankton can quickly regulate PETC throughput as a mechanism to maintain a maximal P^{cell}_{max} in spite of damage to upstream processes (such as the PSII core) [51]. For unknown reasons, other PETC genes did not exhibit strong diel periodicity, including genes encoding plastocyanin (*petE*) and ferredoxin NADP oxidoreductase (*petH*), and the chlorophyll-binding light harvesting complex protein (*pcb*) (Figure 4D).

In general, the diel variation in these photophysiological parameters and the expression of selected genes was consistent with those observed by others for *Prochlorococcus* (PCC 9511) [5,52], but Bruyant et al. (2005) found that the photochemical efficiency of PSII (F_v/F_m) and the absorptional cross section of PSII (σ_{PSII}) varied inversely with light level, while we found little difference in F_v/F_m over the diel cycle. They also observed stronger diel variation in the antenna protein Pcb gene transcript [52] than we did (Figure 4D and Figure S3). We speculate that these differences may be related to the 4-fold lower photon flux used in our study (232 μmol quanta m^{-2} s^{-1} maximum) relative to theirs (912 μmol quanta m^{-2} s^{-1} maximum), perhaps resulting in less stress on the photosystems. Differences in strains used (MED4 versus PCC 9511) may also have played a role.

High-light inducible proteins (HLIPs)

High-light inducible genes encode a family of photosystem associated proteins in cyanobacteria [53,54] that are upregulated in response to environmental perturbations such as nutrient, light and temperature stress [29,30,55,56] and provide a fitness advantage during exposure to high light [55]. They are thought to be involved in the protection of the photosystems from excess light energy although the mechanism for this is under debate [53,54,57,58]. High-light adapted *Prochlorococcus* ecotypes, such as MED4, have over 20 copies of the *hli* genes [3,31,59]. Four of the MED4 *hli* genes are found in almost all marine cyanobacteria in a single copy and their genome context is conserved. In contrast, many of the other *hli* gene types are found in multiple copies in the MED4 genome, are located in genomic islands [31] and are thought to have originated from phages [31,60]. This made us wonder if these two classes of *hli* genes had distinguishable expression patterns under these optimal growth conditions.

All *hli* genes of MED4 were expressed during our experiment, and most of them were periodic (19 out of 22) (Figure 4E, Table S5). Intriguingly, the 4 single copy *hli* genes that are found in all *Prochlorococcus* (i.e. are "core" genes) each have peak expression at a different time of day, spread over the diel cycle (Figure 4E). One of them (*hli1*) has the same expression pattern (Figure 4E) as *psbH* (Figure 4A), which encodes the PSII gene product to which an HLIP binds in *Synechocystis* PCC6803 [53]. The multi-copy *hli* genes also show peak expression at different times of the day (Figure 4F), but not in any way that distinguishes them from the single copy genes. Expression of both single copy (*hli20*) and multi-copy (*hli10*) genes co-varied with P^{cell}_{max} (Figure 4E, F) whereas expression of other *hli* genes however showed no such correlation, peaking at sunrise (*hli1*), sunset (*hli3* and *hli11*), and many even at night (e.g. *hli2* and *hli4*) (Figure 4E, F).

It is striking that there is always at least one *hli* gene upregulated during any four-hour window of the light-dark photoperiod (Figure 4E, F Table S5), suggesting that the different gene products function at discrete stages in the light-dark cycle. They may, for example, serve to keep photosynthetic machinery running near its maximal capacity (i.e. $P^{cell}/P^{cell}_{max} \approx 1$) for a larger portion of the day (Figure 3D). Roles *hli* genes play at night are unknown, but their distributed timing of expression suggests that their activities are more diverse than originally thought. Furthermore, fifteen of the multi-copy *hli* genes that displayed diel variation in expression, but none of the single copy *hli* genes, are also upregulated when MED4 is subjected to environmental stressors [27,29,30]. Thus it appears, for the multi-copy *hli* genes at least, that they play a role in both life of the cell under optimal growth conditions as well as in response to specific environmental stressors, suggesting multiple levels of regulation.

Carbon metabolism and aerobic respiration

Transcripts of genes involved in carbon fixation and storage, carbon catabolism, and respiratory electron transport all show marked diel oscillations, and their timing relative to each other, and the light-dark cycle, offers evidence of tight coordination and phasing of these metabolic pathways. Several reactions are used by multiple pathways that are temporally distinct, presenting a regulatory challenge to the cell. Our probing this phenomenon identified regulatory genes that may be important in orchestrating the flow of carbon and energy in the cell over the course of the light-dark cycle.

Carbon fixation and storage. The entire suite of genes encoding the pathway for carbon fixation and glycogen biosynthesis in MED4 [3] was maximally expressed at dawn (Figure 5 A,B, Table S6). This is consistent with studies of selected genes in this pathway in *Prochlorococcus* [5,8], and is the molecular mechanism initiating the conversion of CO_2 to biomass observed in our physiological analyses (P^{cell}, Figure 3A). Bicarbonate (the predominant source of inorganic carbon in the oceans) is first converted, by carbonic anhydrase (*csoS3*), to carbon dioxide in the carboxysome [61]. This is fixed by Rubisco (*rbcLS*) and proceeds through the Calvin cycle, generating net phosphoglyceraldehyde (PGALD) for anabolism. Some of the fixed carbon is also diverted to biosynthesis of the carbon and energy storage molecule glycogen (via *glgA*, *glgB*, and *glgC*). The anticipatory up-regulation of genes during the dark period is likely responsible for the immediate increase in photosynthetic activity (P^{cell}, Figure 3A) and biomass (Figure 2A) observed once the light energy could be captured.

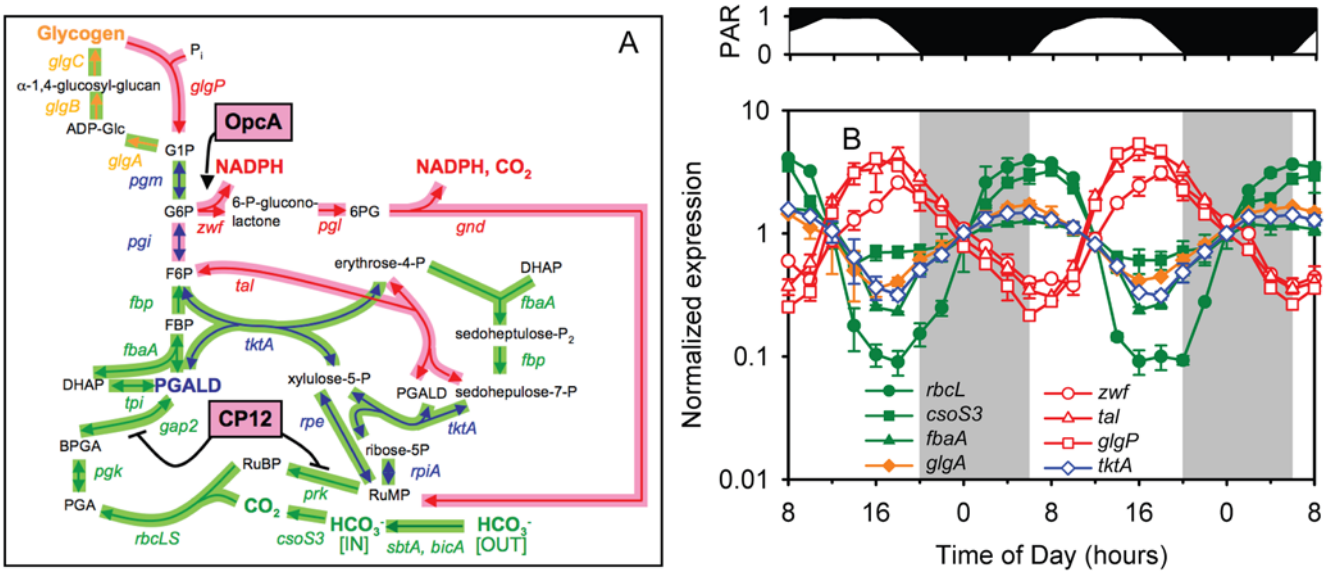


Figure 5. Expression patterns of carbon metabolism genes. (A) Overlay of the Calvin cycle, pentose phosphate pathway, and glycogen metabolism pathways of MED4. Genes and enzymatic reactions (arrows) are color coded by function: green for Calvin cycle, orange for glycogen biosynthesis, red for glycogen catabolism and pentose phosphate pathway, and blue for genes and reactions that are shared between Calvin cycle or glycogen biosynthesis and pentose phosphate pathway. Reactions activated (arrow) or inactivated (dash) by OpcA and CP12 are noted. Peak gene expression was either at sunrise (light green shading of arrows) or sunset (pink shading of arrows and boxes). (B) Expression time course of selected Calvin cycle (green), glycogen biosynthesis (orange), pentose phosphate (red), and dual function (blue) genes. Selected genes are *rbcL*, *csoS3*, *fbaA*, *glgA*, *zwf*, *tal*, *glgP*, and *tktA*. Error bars represent one standard deviation of the mean for the replicate cultures.
doi:10.1371/journal.pone.0005135.g005

Carbon catabolism. Energy for the nighttime activities of *Prochlorococcus* cells (e.g. cell division, nucleotide biosynthesis) likely comes in the form of NADPH, which is generated from the catabolism of stored glycogen via the oxidative pentose phosphate pathway [3] (Figure 5A). Genes for glycogen degradation (e.g. glycogen phosphorylase, *glgP*) and the oxidative pentose phosphate pathway in MED4 had peak expression at sunset (Figure 5A,B, Table S6), thus apparently maximizing their potential for nighttime use of their products, as has been noted for other cyanobacteria [41,43,44]. Notably, the pentose phosphate pathway shares several reactions with the Calvin cycle. As discussed below, *Prochlorococcus* appears to use several mechanisms to regulate these two intersecting pathways.

Respiratory electron transport. In cyanobacteria, respiration occurs in both the cytoplasmic and thylakoid membrane, and in the latter, shares much of the electron-transport machinery with photosynthesis. A key component in both respiration and cyclic photosynthesis is NAD(P)H dehydrogenase. *Prochlorococcus* has two NAD(P)H dehydrogenases, one of the canonical type I (NDH-I) and one of type II (NDH-II) [3]. The latter, composed of a single protein subunit thought to play a regulatory role in other cyanobacteria [62], was not expressed at detectable levels in our study (Table S4). The former are multiprotein complexes, consisting of at least 15 subunits in *Synechocystis* PCC 6803 [62]. MED4 has homologs to all 15 subunits, including 2 paralogs of *ndhD*: PMM0150 and PMM0594. In freshwater cyanobacteria, several complexes of NDH-I exist that contain different paralogs of NdhD and NdhF and that have distinct functions: respiration, cyclic electron transport of photosystem I, and carbon dioxide uptake (the latter is absent in *Prochlorococcus*) [63–65]. Like most of the other *ndh* genes, *ndhD* paralog PMM0150 peaked at sunset, consistent with a role in respiration (Table S4). In sharp contrast, *ndhD* paralog PMM0594 peaked at sunrise (Table S4). These results suggest that MED4 has two different NDH-I complexes: one containing the NdhD encoded by PMM0594 that functions in cyclic electron transport of PSI during the day, and a second one containing the NdhD encoded by PMM0150 for aerobic respiration at night.

In both photosynthesis and respiration, electrons are passed to the plastoquinone pool, then to cytochrome b_6f [41], plastocyanin, and then photosystem I (photosynthesis) or cytochrome *c* oxidase (respiration). Given that cytochrome b_6f is used by both photosynthesis and respiration, it is unclear which process would be favored in a periodic expression pattern. Genes encoding b_6f peaked at mid-day in our experiment (Figure 4C and Table S4), suggesting that the product of this gene is in greater demand for photosynthesis than respiration. We postulate below a similar explanation for other dual-use enzymes that are found in the Calvin cycle and the pentose phosphate pathway. Conversely, for cytochrome *c* oxidase, which is used only in respiration, all subunits had maximal expression at sunset and minimal expression at sunrise (Figure 6 and Table S4). This pattern matches that of glycogen degradation and the pentose phosphate pathway, which presumably supplies NADPH and its electrons for respiration. It remains to be determined what fraction of NADPH from the pentose phosphate pathway is used for respiration and what fraction is used for other processes, such as nucleotide reduction and combating oxidative stress.

ATPase couples the proton gradient, created by electron transport, to ATP synthesis. ATPase should function to generate ATP during both photosynthesis and respiration. It is thus curious that the diel expression pattern of all subunits match those of the Calvin cycle genes (Figure 6, Table S4). This invites the hypothesis that gene expression has been optimized to handle the greatest *demand* for ATP (carbon fixation), rather than the potential greatest *production* of ATP (photosynthetic light reaction and/or respiration).

Dual-use enzymes and intersecting pathways. Metabolic networks often employ the same enzyme, and the chemical

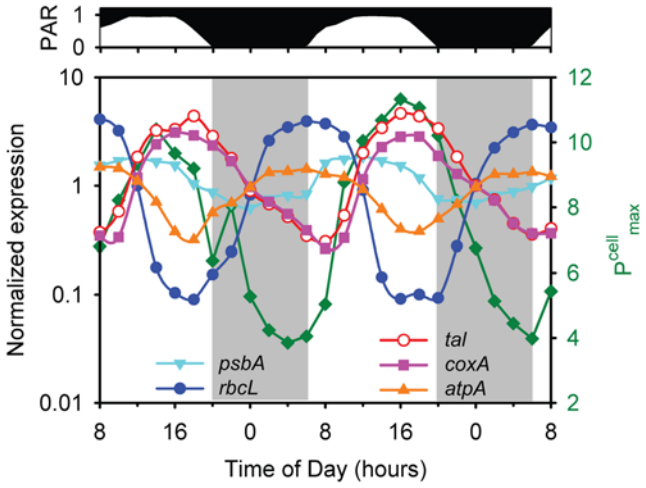


Figure 6. Relationships of carbon and energy metabolism. Expression time course of representative genes involved in photosynthetic electron transport (*psbA*), Calvin cycle (*rbcL*), pentose phosphate pathway (*tal*), respiratory electron transport (*coxA*), and the proton-translocating ATPase (*atpA*) is shown. For comparison, P^{cell}_{max} (see Figure 4B) is also reported (green squares). For clarity, error bars representing sample-to-sample variability in gene expression are not shown.
doi:10.1371/journal.pone.0005135.g006

transformation it catalyzes, for several different purposes. Indeed this is the case in cyanobacteria, particularly vis a vis pathways that are partitioned between night and day. While this could provide efficiency, it presents a problem: If an enzyme is used at day and at night, and expression of its gene is periodic (as are most genes in *Prochlorococcus*), should it be maximally expressed at day or at night?

One of the most striking examples of dual-use enzymes in cyanobacteria is the shared enzymes of the Calvin cycle and the pentose phosphate pathway. Six reactions, catalyzed by five enzymes, are shared between these pathways (Figure 5A, blue arrows). Like most genes in *Prochlorococcus*, the transcripts of these exhibit a diel periodicity: all with peak expression at sunrise and minimal expression at sunset, similar to the expression of the Calvin cycle genes. Assuming a tight coupling between the timing of gene expression and protein levels, this may seem counterproductive for nighttime respiration. It is possible, however, that flux through these reactions is more intense when the Calvin cycle is operating, so greater quantities of these enzymes (and hence transcripts) are needed during the day. Additionally, we note that four of the six shared daytime-maximal reactions run in opposite directions in the two pathways. Intriguingly, work in other organisms has shown that the equilibrium constants for these four reactions all favor the pentose phosphate pathway direction: ribulose-5-P isomerase (*rpiA*), ribulose-5-P epimerase (*rpe*), and both transketolase (*tktA*) reactions [66]. Given these equilibria, particularly that of the final transketolase reaction ($K_{eq} = 17$), it is plausible that smaller quantities of enzymes at night are sufficient to yield significant flux through the pentose phosphate pathway, and that the diel periodicity of the transcripts that encode these enzymes serves to keep these channels equally open for both the Calvin cycle and the pentose phosphate pathway.

The alternation in carbon flow between these two pathways raises another issue. How does the cell direct carbon in the required direction and mediate the oscillation between these two pathways? Perhaps translational control is sufficient; abundance of enzymes exclusive to one pathway could steer the overall flux in the required direction. It appears, however, that additional regulatory mechanisms, involving the post-translational regulatory proteins, CP12 and OpcA, are operating to control the switch between Calvin cycle and pentose phosphate pathway. CP12, an intrinsically unstructured protein, has been shown to directly inhibit the Calvin cycle during nighttime conditions in both cyanobacteria and green plants [67,68]. In our data set, the gene encoding CP12 (PMM0220) is maximally expressed in the evening (Figure 5A). The oxidizing conditions of a cyanobacterial cell at night are known to trigger CP12 to bind and deactivate the Calvin cycle enzymes phosphoribulokinase (PRK, *prk*), and glyceraldehyde-3-P dehydrogenase (GAPDH, *gap2*) [67]. Thus, nighttime expression of CP12 likely serves to shut off key steps in the Calvin cycle of MED4 to let the pentose phosphate pathway proceed unhindered. At the same time that the gene encoding CP12 is induced, so is that for OpcA (Figure 5A), an allosteric effector of the first enzyme of the oxidative pentose phosphate pathway, glucose-6-P dehydrogenase (G6PDH, *zwf*). OpcA is known to increase G6PDH affinity for glucose-6-P more than 100-fold in other cyanobacteria [69]. From these expression patterns, we suggest that at the same time CP12 restricts carbon flow through the Calvin cycle, OpcA appears to redirect carbon flow through the pentose phosphate pathway. Together with the alternate phasing of expression of the genes encoding the pathway enzymes, induction of regulatory proteins that activate or deactivate key enzymes of the two pathways may be the crucial events that facilitate the temporal separation of the Calvin cycle and pentose phosphate pathway.

Offset of transcripts for photosynthetic light and dark reactions. We found an interesting difference in phasing between expression of the Calvin cycle genes and those that encode the light reaction of photosynthesis, which provide the Calvin cycle with energy and reducing power (Figure 6). Most of the genes of the photosynthetic electron transport chain reach peak expression levels in the middle of the light period, where as Calvin cycle genes, such as *rbcL*, had peak expression levels at dawn and transcript levels were minimal toward the end of the day. This may account for the slight uncoupling of P^{cell}_{max} and α^{cell}_{max} (Figure 3B): maximal carbon fixation via the Calvin cycle may occur before maximal light utilization due to the offset in timing of the synthesis of the proteins involved (Figure 6).

Assuming our inference from transcript levels is correct, why does expression of Calvin cycle genes precede expression of the light reaction genes? Perhaps it allows the cell to take immediate advantage of reducing power at sunrise, thus minimizing the dependency of photosystem usage during periods of high (and damaging) light intensity later in the day. Additionally, significant down-regulation of *rbcLS* in the afternoon may help limit the amount of photorespiration (i.e. oxygenation of ribulose-1,5-bisphosphate, rather than carboxylation) during periods of high light and O_2 production. All of these interpretations remain hypotheses until they can be explored at the protein and metabolome levels rather than transcript level alone.

In summary, over the course of the photocycle, the energy source for *Prochlorococcus* undergoes dramatic variation. The amount of light available at a particular time determines the source of electrons that are used for NADPH production— either water or glycogen. Our transcriptome analysis has generated hypotheses about how the transitions to the different modes of energy and carbon metabolism are mediated at the level of gene expression. Genes of the oxidative pentose phosphate pathway and the respiratory electron transport chain, which together turn glycogen into NADPH and then ATP, cycle with sunset maxima and sunrise minima — 180° out of phase with those of the Calvin cycle (Figure 5B, Figure 6). These patterns, plus those of the

regulatory proteins CP12 and OpcA (see above), suggest how the *Prochlorococcus* cell transitions from daytime photosynthetic carbon fixation to the nighttime shutdown of the Calvin cycle and induction of the respiratory pathway, which likely accounts for the observed nighttime decline in photosynthetic capacity (P^{cell}_{max}) (Figure 3B).

Diel periodicity of nutrient acquisition and assimilation

Nutrient transporters are a critical link between the cell and its environment. One might predict *a priori* that the transporters for carbon, phosphorus, and nitrogen are maximally expressed near the time of greatest demand by the cell each element. In the mildly-alkaline oceans, the vast majority of inorganic carbon is in the form of bicarbonate, and we might expect demand to be highest during the day, when cellular biomass increases (Figure 2A). MED4 contains homologs to two sodium-bicarbonate symporters, *sbtA* and *bicA* [63], that are likely in the same operon. As predicted from supply and demand considerations, expression of both genes cycled synchronously with the Calvin cycle and carboxysome genes (Figures 5A, 7A), sharing the same cluster (16) with most of them (Table S6).

P-limitation exerts a strong selective force on the composition of *Prochlorococcus* and its genome [28], as evidenced by the fact that most of cellular P is in DNA and RNA [70], and essentially none in phospholipids [71]. We expect the cell's greatest demand for P to be during the day-to-night transition, i.e. the period of DNA replication (Figure 2D), high total mRNA accumulation (Figure 1), and peak expression of the nucleotide biosynthesis genes (Table S1). With the greatest demand for P during the evening, the expectation was that peak expression of phosphate uptake genes would also be in the evening. Indeed, this was the case for the trans-membrane and ATP-binding cassette components of the ABC-type phosphate transporter (together the *pstCAB* operon)(Figure 7A). In contrast, *pstS*, which encodes the periplasmic phosphate binding protein component of the transporter has high transcript levels with very weak periodicity with a late-night maximum (Figure 7A). We speculate that this near-aperiodic expression serves to maintain a constant (high) concentration of PstS in the periplasm, and trap any phosphate that may enter throughout the photoperiod, while transport *per se* is maximized at night, at the time of highest demand.

Previous studies have shown that alkaline phosphatase (encoded by *phoA*), which cleaves phosphate from organic sources in the periplasm, and an alkaline phosphatase-like protein of unknown function (encoded by *dedA*) exhibit contrasting expression patterns in MED4: *phoA* is highly upregulated under P-starvation while *dedA* is not [28,72]. In this study we observed measurable *dedA* expression with a maximum in the mid-afternoon, while *phoA* displayed periodicity similar to that of phosphate uptake genes, with peak expression just after dark (Table S7), just anticipating the greatest cellular demand for P. We postulate that *dedA* may be responsible for the low constitutive alkaline phosphatase activity (APA) that has been documented for MED4, and other *Prochlorococcus* strains that lack *phoA* [72]. Measuring APA over a diel cycle under both replete and P-starvation conditions will help tease apart the roles of these and other phosphatases.

The cellular demand for nitrogen is more complex than that for phosphorus, with the majority of nitrogen contained in proteins in addition to nucleic acids. Nitrogen demand for protein synthesis is likely to closely follow that for mRNA which shows a bimodal pattern of expression at both sunrise and sunset (Figure 1), but is also distributed at moderate levels throughout the day. Given this complexity, it is difficult to postulate any sort of supply/demand relationship for this element without proteomic data. Thus we simply offer some selected observations regarding the periodicity of expression, or lack thereof, of N-related genes.

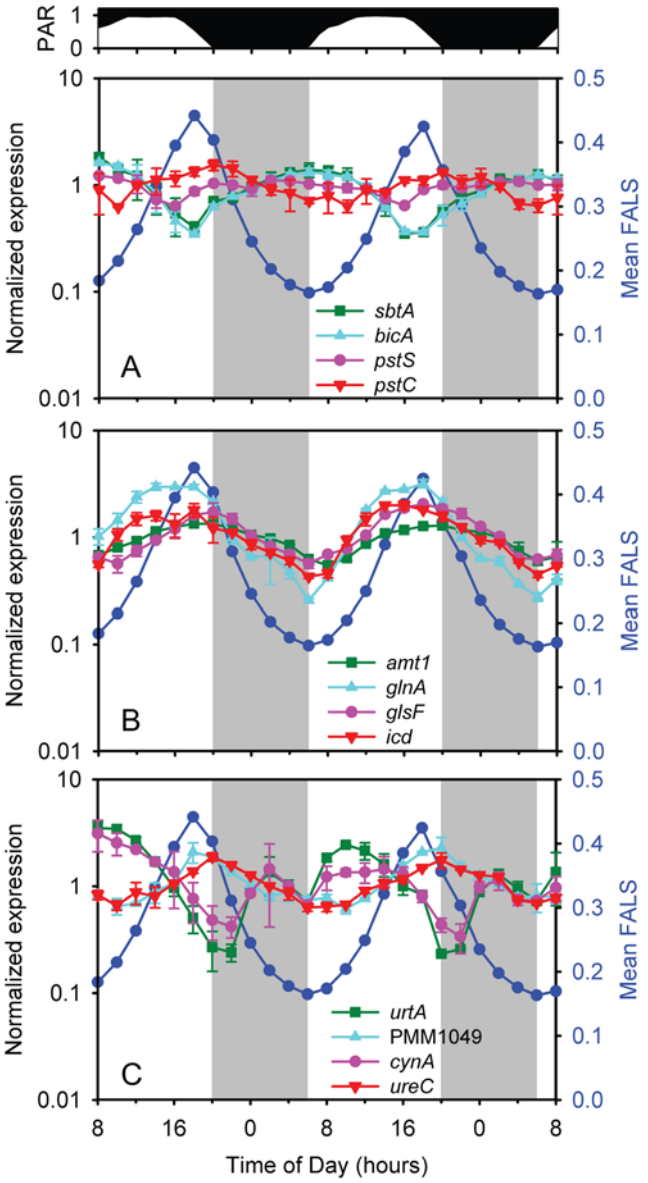


Figure 7. Expression time course of nutrient transport and assimilation genes. (A) Transporter genes for bicarbonate (*sbtA*, *bicA)* and phosphate (*pstS*, *pstC*); (B) ammonium transport (*amt1*) and assimilation (*glnA*, *glsF*, *icd*); and (c) transporter genes for urea (*urtA*), oligopeptides (PMM1049), cyanate (*cynA*), and for a urease subunit (*ureC*) are shown. Cell size (FALS) variation over the photocycle is also presented (dark blue circles) for comparison. Error bars represent one standard deviation of the mean for the replicate cultures.
doi:10.1371/journal.pone.0005135.g007

Amino acid synthesis genes show bimodal patterns of expression (Table S1) with peaks at sunrise and sunset, while expression of ribosomal (protein) genes is highest throughout the night. We note further that the expression of different genes in the synthesis pathway for the same amino acid peak at different times over the cycle suggesting complex patterns in N-demand for protein synthesis. Transcription of the ammonium transport gene (*amt1*) peaked in the evening, near sunset (Figure 7B). Transcript levels of *amt1* were at least an order of magnitude higher than for other nitrogen metabolism related genes (data not shown), and displayed

low cyclic amplitude over the diel cycle (Figure 7B). This suggests that, similar to the phosphate periplasmic binding protein encoded by *pstS*, there is a need for constant high expression of Amt1 to ensure efficient scavenging of any available ammonium from the nutrient deplete waters *Prochlorococcus* inhabits in nature. The ammonium assimilation pathway genes exhibited the same periodicity as the transporter (Figure 7B), peaking in the evening. Ammonium is assimilated into organic compounds via the glutamine synthetase (GS) – glutamate synthase (GOGAT) pathway (encoded by *glnA* and *glsF* respectively) and the carbon skeleton for its incorporation is 2-oxoglutarate (2-OG) [73]. The ammonium assimilation pathway genes exhibited the same periodicity as the transporter (Figure 7B), peaking in the evening. 2-OG is produced from isocitrate by isocitrate dehydrogenase (*icd*) which was also maximally transcribed in the evening (Figure 7B). This suggests that the major source of the carbon skeleton (2-OG) for ammonium assimilation is not phosphoglyceraldehyde generated directly from photosynthesis, but rather from glycogen stores.

Although ammonium is preferred [73], *Prochlorococcus* can utilize different sources of nitrogen for growth. Urea and cyanate can serve as nitrogen sources in *Prochlorococcus* [30,74]. Although they were not in the media during this experiment, their transporters (*urtAB* and *cynA*) were expressed with a complex pattern of transcription with maxima soon after sunrise and a secondary peak at night (Figure 7C). Both of these genes encode ABC-type transporters and their peak expression coincides with that of the ATPase gene (Figure 6). Urease genes (which convert urea to ammonium) had maximal expression in the evening (Table S8), consistent with the timing of expression of ammonium transport and assimilation genes but different to that for the urea transporter genes. Finally, recent data suggests that *Prochlorococcus* can take up methionine and leucine, and that their accumulation is significantly higher at dusk than at dawn [75]. This matches the timing of expression of ammonia uptake and assimilation genes, as well as that for the predicted oligopeptide permease gene (PMM1049) (Figure 7C) in our experiment.

Previously reported P and N starvation responses in *Prochlorococcus* MED4 revealed the up-regulation of many genes besides those directly involved in uptake and assimilation [28,30]. We examined the periodicity of these genes in our experiment and found that their behavior fell into two distinct subsets: some had transcription patterns similar to phosphorus and nitrogen assimilation genes (data not shown), and some were not expressed above background at any time point. Genes in the first group therefore appear to be subjected to multiple layers of regulation. Their induction during nutrient starvation indicates a role in stress response while their diel oscillation suggests that they also play integral roles in nutrient assimilation even in cells grown under optimal conditions. Genes in the latter group, however, may be stress-response specific being highly induced from background levels during nutrient starvation. These include PMM1403 and PMM0721, genes of unknown function which are upregulated during P-starvation [28], and the nitrogen transcriptional activator *ntcA* and PMM0958 a gene of unknown function which is the most highly upregulated gene during N-deprivation [30].

Regulation

The mechanisms that regulate and choreograph the cyclic gene expression patterns we have described are yet to be unveiled. It is likely, however, that they involve (1) transmission of light as a signal for gene expression through a photoreceptor-regulatory pathway and (2) a diel oscillator of some sort. While the expression of some genes, such as *psbA*, varies in direct proportion to available light, this is not true for the majority of periodic genes. In particular, if light is the sole trigger for up or down regulation, it is difficult to reconcile this with the night-time induction of genes such as *rbcL* that seem to anticipate the coming of dawn. Hence the most likely "master controller" of the transcriptome would appear to be some sort of endogenous oscillator like a circadian clock. Yet as discussed above, *Prochlorococcus* lacks key components of the cyanobacterial clock, such as *kaiA* and *cikA* and does not display cyclic gene expression under constant conditions [12]. Are there any clues as to regulation in the patterns of expression of the known clock genes in *Prochlorococcus*?

The *kaiB* clock gene of MED4 exhibited strong diel periodicity in our experiment, with a maximum at dawn and a minimum at sunset (Figure 8A). *kaiC* showed low, albeit significant diel periodicity, peaking near the onset of darkness. While the *kaiB* pattern resembles that reported by Holtzendorff et al. (2008) for *Prochlorococcus* PCC 9511, the *kaiC* pattern is just the opposite: they found that *kaiC* peaked at dawn, in phase with *kaiB* (albeit with a small secondary peak just after dark, in phase with the peak we observed). This difference is puzzling. At this time, we can only add that we found the same weak periodicity of *kaiC*— with a small peak after the onset of darkness — in our Pilot Study (see methods) using quantitative reverse transcription PCR (Figure S3), as was observed in our study with the arrays.

MED4 has homologs to *sasA* and *rpaA*, which in another cyanobacterium encode the histidine kinase and cognate response regulator that are essential for transmission of the clock's output to the genome [23,76]. Transcripts of the *sasA* homolog oscillate in a pattern almost identical to *kaiC* in our experiment (Table S9), while the *rpaA* homolog peaks twice per 24-hour period, at 10:00 and 12 hours later at 22:00 (Figure 8A). *cpmA* is also involved in clock output [77], and the homolog of this gene in MED4 shows weak but significant periodicity, peaking just prior to *kaiB* at night (Figure 8A). Hence, all of the homologs of circadian clock-related genes that MED4 possesses exhibit diel periodicity in transcript abundance. The significance of these periodicities is currently unknown, given that SasA protein accumulation in *Synechcoccus elongatus* PCC 7942 was constitutive over a light-dark photocycle [23], and that two-component regulatory systems such as SasA-RpaA are themselves regulated primarily at the post-translational (phosphorylation state) level.

All five sigma factors in MED4 (one group I plus four group II) cycle with unique phase relations to the diel photocycle (Figure 8B). Holtzendorff el al. (2008) reported the different periodicities of two of the sigma factors (PMM1629 and PMM1697), and here we confirm and extend those results to include all five. PMM1289, PMM0577, and PMM1697, cycle similarly but not identically: PMM1289 begins to accumulate in early morning, peaks at midday, and is followed by PMM0577 and PMM1697 with a 2 hour offset (Figure 8B). The predicted principal (group I) sigma factor (*rpoD*, PMM0496) peaks two hours after the onset of darkness, and the final sigma factor (PMM1629) peaks at dawn. PMM1629 has the same phasing as *kaiB*, which raises the possibility that the former regulates expression of the latter, although the reverse scenario might also be true. In addition, PMM1629 is also in expression cluster 15, which is significantly enriched with Calvin cycle and ATPase genes (Table 1), raising the possibility that it controls expression of a photosynthesis regulon.

We suspect that differential phasing of expression of the sigma factors may contribute significantly to the diel expression patterns of the rest of the transcriptome. In other cyanobacteria, the inactivation of group II sigma factors can cause defects in *psbAI* and *kaiB* circadian expression [78,79], and they are thus thought be involved in transducing the clock output signal to the genome: While group II genes are transcribed in phase (in contrast to

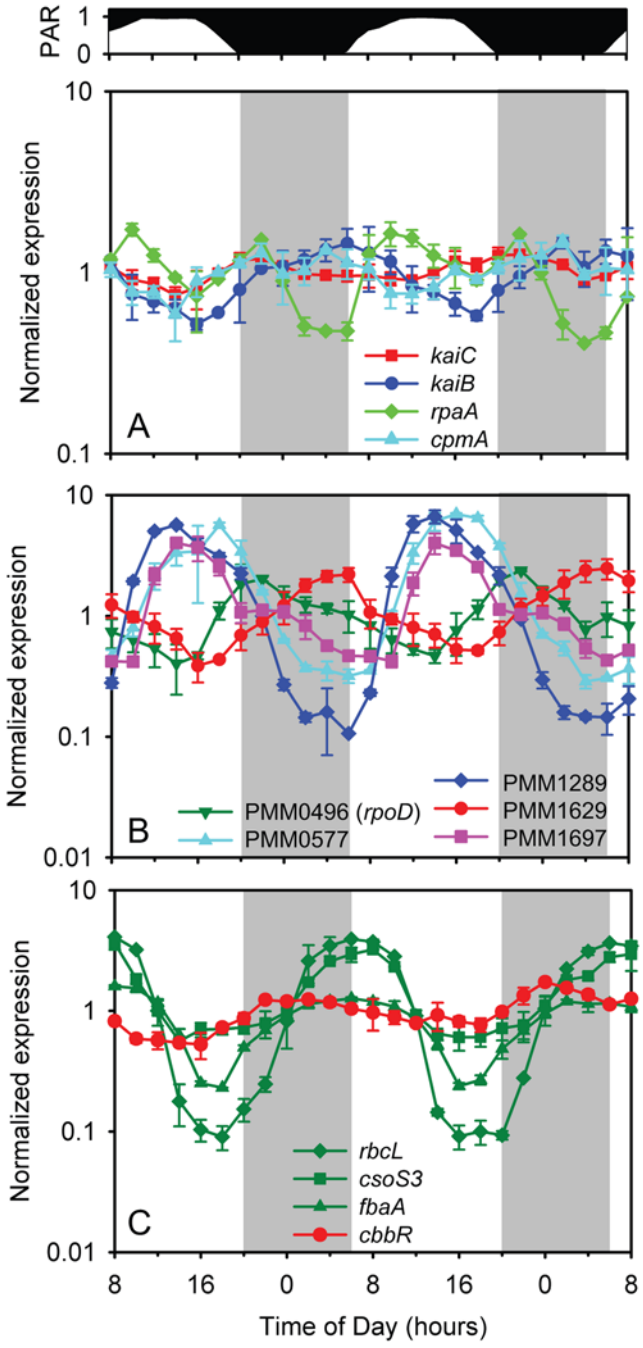


Figure 8. Expression time course of regulatory proteins. (A) Circadian clock genes *kaiB* and *kaiC*, and clock output genes *rpaA* and *cpmA*; (B) Sigma factor genes PMM0496 (*rpoD*), PMM0577, PMM1289, PMM1629, and PMM1697; and (C) *cbbR*, and representative Calvin cycle genes *rbcL*, *csoS3*, and *fbaA* are shown. Error bars represent one standard deviation of the mean for the replicate cultures.
doi:10.1371/journal.pone.0005135.g008

MED4), they are thought to confer the observed staggered variation in expression of target genes through differential phasing of activity [78]. It is argued that the sigma factors all compete for the core RNA polymerase, but are activated (perhaps via translation) at different times in the photoperiod, thus with their different affinities for both core polymerase and the suite of promoters within the genome, effectively turn different sets of genes on and off at different times [78]. In support of this hypothesis, cross-regulation between the group II sigma factors during the day and during transitions from dark to light has been reported in *Synechocystis* PCC 6803 [80,81], and one (SigB) regulates transcription primarily in the dark, while the another (SigD) does so primarily in the light [82].

The transcripts of most (46 of 54) of the other putative transcription factors [3] cycle over the light – dark cycle; only five of those expressed were aperiodic, and only four were below the signal threshold to be considered expressed (Figure 8, Table S9). Thus a large fraction of the transcription factors appear to be active during optimal growth under a light-dark cycle. Might the diel periodicities of these regulators establish the periodicities of other genes over the photocycle?

There are two examples that invite investigation in this regard — one involving carbon acquisition and the other nitrogen acquisition. The expression of the Calvin cycle genes is concurrent with, or proceeded by, the increased expression of PMM0147 (Figure 8C). This is a homolog of the regulatory protein gene *cbbR*, paralogs of which are known to play roles in carbon metabolism in *Synechocystis* PCC 6803 [83]. The MED4 gene PMM0147 is most similar to the *lysR* paralog, which is essential in *Synechocystis* PCC 6803, and believed to regulate *rbcLS* in this system. As the MED4 genome lacks any other CbbR-type paralogs, it appears that PMM0147 is a functional ortholog of *lysR*, and its expression pattern further implicates it in the diel regulation of the Calvin cycle genes, including *rbcLS*. Regarding N-acquisition, PipX is a transcriptional co-activator that is required for NtcA-dependent transcription of nitrogen metabolism and transport genes under N-stress [84]. The *pipX* gene in MED4 displayed maximal expression at sunset (Table S9) as did most of the N metabolism genes. It is worth exploring whether the diel periodicity of nitrogen transport and assimilation genes that peak in the evening in *Prochlorococcus* is mediated by the periodicity of *pipX*.

Comparison of the light-dark entrained transcriptome with that from cells grown in continuous light

Given the tight alignment of the *Prochlorococcus* cell cycle to the light-dark cycle, and the choreographed gene expression and physiology, it is perhaps surprising that *Prochlorococcus* can be maintained at maximal growth rates under continuous-light [85]. To begin to understand the adjustments in cellular physiology that enable growth under continuous light, we compared our diel transcriptional profiles to the continuous light profiles (of nutrient-replete control treatments) from previous experiments with MED4 [28,30]. If growth in continuous light simply effects the complete desynchronization of the population that has been shown to occur in cultures shifted from light dark cycles to continuous light [12], then we would expect gene expression in continuous light to represent some average expression over all time points of the diel cycle. Instead, we found 39 genes whose RMA-normalized expression is significantly higher (>2-fold, $q<0.05$) in continuous light compared to the mean over the diel cycle, and 17 genes significantly lower (>2 fold, $q<0.05$) (Table S10). The products of the former 39 genes include ten high light induced proteins, 3 group II sigma factors (PMM0577, PMM1289, and PMM1629), and five heat shock proteins, including GroEL/GroES. While interpretation of this comparison cannot be conclusive because the integrated 24 hour photon flux in the continuous light experiments was 35–60% that in the diel experiment, and the cells were growing at lower growth rates, it is striking that "stress-related" genes constitute a significant fraction of the overrepresented transcripts under continuous light conditions. A side-by-side comparison of the continuous light and diel transcriptome of cells grown under the same daily integrated photon flux would be

instructive, as would a comparison of the transcriptome of cells maintained at the same growth rate, in continuous light and on a light dark cycle. Combined, these experiments would help bring to light the cell's response to continuous illumination at the molecular level.

General conclusions and future directions

This study brings us one step closer to the broad goal of developing *Prochlorococcus* as a model for integrative systems biology — i.e. to understand its cellular architecture, variability, and the forces that shape the *Prochlorococcus* meta-population in the global oceans. A description of the diel transcriptome of the cell in the context of its photophysiology and cell cycle is essential for the development of metabolic models of the cell. Coupled to future proteomics and metabolomics studies, we will have a more complete understanding of how diel gene expression, and the timing of protein activity, is controlled at the level of transcription, translation, and post-translational regulation. Furthermore, this data set is an invaluable reference for interpreting the growing open-ocean meta-transcriptomics database, in which *Prochlorococcus* transcripts are highly represented [86].

The *Prochlorococcus* system is particularly useful for this type of study because of the tight synchrony of the cells when grown on a light-dark cycle, ensuring that the gene expression patterns reflect what one would measure in an individual cell as it progresses through its cell cycle. This schedule of events appears highly choreographed and aligned with the photocycle. The cell is 'born' sometime during the dark period and by the time dawn arrives, the transcripts of the full complement of Calvin cycle and carbon concentrating mechanism genes are maximally abundant, as well as those of many genes encoding members of the photosynthetic electron transport chain. This primes the cell for photosynthesis and net biomass accumulation, which begins as soon as light hits the cell. Expression of some other photosynthetic genes, such as *psbA*, appears to be under a different regulatory regime as they directly track light intensity, peaking at noon. In other cyanobacteria, *psbA* expression is controlled by light (and/or redox state) and the circadian clock [49,87,88], and this may be the case for MED4 as well. As dusk approaches, expression of DNA polymerase and other genes involved in DNA synthesis are maximally expressed, closely followed by the onset of chromosome replication (S phase) of the cell. As day transitions to night, genes encoding the divisome become maximally expressed, and the cell undergoes cell division sometime during the night, completing the cycle. The day to night transition is also marked by a switch in energy metabolism from photosynthesis to aerobic respiration, and in carbon metabolism from CO_2 fixation to catabolism of glycogen, both of which are manifested in the changes in gene expression.

The robust periodicities of gene expression in *Prochlorococcus* suggest strong selection for the coordination of cellular processes in face of the oscillating energy supply. Indeed, relative fitness of *Synechococcus elongatus* PCC 7942 clock mutants has been shown experimentally to be a function of how closely their endogenous period matches that of the environmental light-dark cycle [89]. While we do not have similar direct evidence for *Prochlorococcus*, the temporal partitioning of the expression of Calvin cycle and Pentose Phosphate Pathway genes (Figure 5), for example, suggests that selection under the daily photocycle has shaped these patterns. These two pathways play opposite roles in the cell — the former trades energy for fixed carbon, and the latter does the reverse — yet they share several enzymes. This would pose a significant regulatory challenge for the cell if both were operating at the same time — a challenge that would be exacerbated by the streamlined regulatory system of this cell.

Gene inactivation (which is currently not possible in *Prochlorococcus*), proteomics, and studies that vary the growth rate (see below) should provide valuable tests of the hypotheses about regulation generated by these descriptive data. In this study, the doubling time (approximately 1 day) matched the 24 hour photoperiod. But we know that the length of the DNA synthesis phase (S) is growth rate independent in *Prochlorococcus*, while the pre-and post-synthesis phases expand with generation time [14]. Thus by varying average cell generation time to offset it from the 24-hour photoperiod, one may be able to see which processes are set by photoperiod, and which by growth rate. It would also be informative to study these diel transcription patterns under nutrient limited conditions. One could then ask questions such as: Does oscillation in the availability of a limiting nutrient influence the choreography of the transcriptome in response to the photocycle?

In the oligotrophic ocean, where seasonality is typically weak and conditions generally change slowly, the diel light-dark cycle is one of the principal features governing temporal variation in microbial community function. *Prochlorococcus* is one of the few, if not currently the only, microbe whose transcripts are represented in relatively high abundance in meta-transcriptomics data from the open ocean [86]. Thus this laboratory study of the tempo of expression in *Prochlorococcus* cultures, compared with data from the field, can help inform the design of oceanographic sampling strategies. The strongest contrast in gene expression levels in our study was not, as might have been expected *a priori*, between midday and midnight, but rather between sunrise and sunset. In fact, expression levels of most genes were equivalent mid-day and mid-night, with some on the upswing and some on the downswing. Hence if resources are limited and one cannot resolve the entire light-dark cycle it would be most important to sample and sequence around dawn and dusk to capture the metabolic pulse of a cell like *Prochlorococcus*. As a dominant primary producer in these systems, this pulse may be important in driving that of other organisms in the microbial food web.

Interactions between *Prochlorococcus*, phages that infect them, their protozoan predators, and competing microbes, are all likely influenced by diel cycling, as well as other environmental factors which may in turn influence their responses to the oscillating energy input. While the complexity of the interactions is daunting, we are beginning to develop tools that bring it closer into focus. Novel ocean ecosystem models are under development that begin to embrace the diversity of metabolic possibility among microbes [90], and we are getting closer to cellular systems models of ocean microbes, in part through studies such as this one. Our hope is that in time, these two types of systems biology models will meet in the middle, such that the interactions between the environment and the cell can be explored at multiple levels of organization, from the genome to the ecosystem. This will open new vistas for understanding the nature, evolution, and regulation of microbial processes.

Materials and Methods

Pilot Studies

Before executing the comprehensive transcriptome analysis using micro-arrays, a pilot study was conducted on axenic MED4 to determine optimal sampling strategies. Quantitative reverse transcription PCR (QRTPCR) was used to analyze the transcript levels of key genes involved in cell cycle processes (*ftsZ*, *dnaA*), photosynthesis (*psbA*, *pcb*, *rbcL*), the circadian clock (*kaiC*) as well as

transcription (*rpoD*). For this study cultures were grown as described below, except on a 12 hour light (approximately 300 μmol Q m^{-2} s^{-1}), 12 hour dark cycle (without dawn and dusk). The growth rate of the culture was 0.47 day^{-1}. QRTPCR was carried out according to the methods described in [91]. The *rnpB* housekeeping gene was used to normalize RNA between samples. The primers used are shown in Table. S11. Transcript levels of the genes analyzed are shown in Figure S3. We present these results here to show that transcript periodicity patterns of these genes are similar to those determined with the arrays in the actual experiment, even though the culture growth conditions were not identical. For the actual experiment conducted for the arrays, the L:D cycle was changed to 14:10, with a dusk and dawn simulation (see below), so the cultures would grow at exactly one doubling per day.

Culture conditions

Axenic strain MED4 was grown in Sargasso Seawater-based Pro99 medium, which provides nitrogen as ammonia and phosphorus as inorganic phosphate. The Pro99 medium was supplemented with 10 mM HEPES buffer (pH 7.5) to maintain pH and prevent CO_2 limitation [85]. Replicate batch cultures were grown in 10 L volumes within 13.25 L acid-washed glass vessels with slow stirring, at 24±0.2°C. This light level provided maximal growth rate for MED4 under the conditions provided (data not shown). Incubations were performed in a modified Percival Scientific (Boone, IA) I-35LL plant growth chamber. Standard 20 W bulbs and supporting ballasts were replaced with 54 W high-output bulbs and supporting ballasts. Creation of a control device allowed for the voltage-regulated variation in light output from these bulbs. This lighting system was programmed to provide a 14 hour light, 10 hour dark cycle, with a gradual increase or decrease of light at experimental sunrise or sunset, respectively. Sunrise initiated at experimental 06:00, ending at 10:00, and sunset initiated at experimental 16:00, ending at 20:00. Maximum light intensity, at experimental 10:00–14:00 was approximately 232 μmol Q m^{-2} s^{-1}.

Every two hours over the 50 hour experiment, 300 mL of the cultures were transferred to centrifuge bottles. Sampling at experimental night time points was performed under very low (<1 μmol Q m^{-2} s^{-1}) red light conditions. Cells were pelleted by centrifugation at 10,000 RPM at 20°C, and resuspended in 1 mL RNA resuspension buffer (200 mM sucrose, 10 mM sodium acetate, 5 mM EDTA, pH 5.2) [29,91]. Samples were snap-frozen in liquid nitrogen and stored at −80°C until processing. At each time point, 3 1 mL aliquots were also prepared for flow cytometry following [92]. To these aliquots, a 0.125% final concentration of TEM grade glutaraldehyde (Tousimis) was added, and after a 10 minute incubation in the dark, these fixed cells were snap frozen and stored in liquid nitrogen.

RNA isolation and quantification

Total RNA was extracted, purified from DNA, and concentrated following Lindell et al. (2005). For microarray analysis, 2 μg of total RNA was labeled and hybridized to the custom MD4-9313 Affymetrix GeneChips®, following standard protocols [29,91]. Raw data were normalized by the Robust Multichip Average (RMA) algorithm [93], via the GeneSpring GX 7.3.1 software (Agilent Technologies).

Flow cytometry and cell cycle analysis

Thawed samples were stained with the DNA stain Hoechst 33342 (0.5 mg ml−1 final concentration) and held at room temperature in the dark for 1 hr prior to analysis following [94,95]. *Prochlorococcus* were enumerated using a modified EPICS V (Coulter) flow cytometer following [96,97]. Relative DNA and chlorophyll concentrations were determined using cellular blue and red fluorescence, respectively, normalized to 0.46 μm carboxylate and 0.47 μm YG bead standards (Polysciences), respectively, following [94]. Cell-cycle parameters were determined using FlowJo cell-cycle analysis software v (TreeStar) from DNA histograms and following [98]. No heterotrophic bacteria (i.e. populations without red fluorescence) were detected over the course of the experiment.

Photophysiology

Photosynthesis irradiance (P-E) curves were measured using the C-14 technique with a conventional photosynthetron [99] as previously described [100,101]. Briefly, 13 1 ml samples were each inoculated with ~0.37 MBq $H^{14}CO_3$, incubated at different light levels in a custom-built, temperature-regulated photosynthetron and terminated after 1 hr with 1N HCl, final concentration. Carbon uptake was quantified using liquid scintillation counting following Barber et al. (1996) [102]. A standard P-E model [103] was optimized to data using a custom written routine following [104] to determine key parameters of photosynthesis, including the light utilization index (α), maximal photosynthesis (P_{max}) and light saturation index (E_k) of the P-E curves as defined by Sakshaug et al. [105]. Rates of photosynthesis for each 2 hr time period in each replicate culture were measured similarly in duplicate except that samples were incubated at ambient light levels with the culture. Single turnover fluorescence induction curves were measured using a Background Irradiance Gradient – Single Turnover fluorometer (BIG-STf) to measure the photosynthetic conversion efficiency (Fv/Fm) and functional absorption cross section (σ_{PSII}) of photosystem II (PSII) as a function of background light intensity as previously described [106]. Duplicate samples from duplicate cultures were dark acclimated for >15 mins, after which single turnover fluorescence induction curves were measured over a range of background light levels. Photosynthetic parameters (Fv/Fm and σ_{PSII}) were estimated by fitting standard models to data to determine values of *Fo* (initial fluorescence), *Fm* (maximal fluorescence), *Fv* (*Fm-Fo*), σ_{PSII} (functional cross-sectional area of PSII) and *p* (PSII connectivity parameter) [107].

Normalization and computational analysis of Affymetrix arrays

Signal intensities for Affymetrix probe sets were calculated and normalized using the Robust Multi-Array Average (RMA) procedure as implemented in the Bioconductor package *affy* [108]. Additionally, we applied the Microarray Suite (MAS 5.0) and "Golden Spike" normalization schemes to study the influence of the chosen normalization procedure [109]. Although some variation in the calculated signal intensities was observed, the main results of the computational analysis remained unaffected.

The detection of periodic expression was based on Fourier analysis, as a recent comparison showed its superior performance compared to other approaches [110]. After averaging over the corresponding time points in both experimental runs, a Fourier score was calculated for the temporal expression pattern of each gene. The Fourier score is defined as

$$F[\mathbf{x}] = \sqrt{\left(\sum_i (\cos(2\cdot\pi\cdot t_i/T)\cdot x_i\right)^2 + \left(\sum_i (\sin(2\cdot\pi\cdot t_i/T)\cdot x_i\right)^2}$$

where $\boldsymbol{x}$ is the standardized expression vector (mean($\boldsymbol{x}$) = 0;

$sd(\boldsymbol{x}) = 1$) for the gene, T is the period (in our case 24 h), and x_i is the measured expression at time point t_i.

To assess the significance of the score obtained, the probability of how frequently such a score would be observed by chance has to be calculated. Thus, a background model for the Fourier score F was generated by fitting autoregressive processes of the order 1 (AR (1)) to the observed time courses and subsequent calculation of F for the generated random expression vectors. Note that the AR(1)-based background models give an improved estimation of the significance of periodic microarray data compared to conventionally used background models based on random permutation [111]. Next, the significance of the measured periodicities was obtained by comparison with the generated background distribution. For each score, a FDR (False Discovery Rate) was calculated representing the fraction of estimated false positives. A FDR-value of 0.10 would indicate that a score larger or equal to the measured one was observed in one out of ten random time courses. This distribution of Fourier scores for measured and generated random time series can be seen in Figure S4. Our model, which implies one peak per period, accounted for the vast majority of periodicity patterns. However, in rare instances, such as *rpaA* (Figure 8A), two major peaks per 24-hour period were observed, and these were usually reported as aperiodic (FDR>0.10) (e.g. *rpaA*, Table S9). Future analyses on this small subset of the genome would validate the periodicity of this interesting category of genes.

Time of peak RNA abundance was determined by two methods. The first consisted of simply identifying the sampling time point where expression was maximal during day 1 and 2. Subsequent averaging the time points leads to the peak time with a resolution of an hour. Considering the distribution derived for all probe sets, a bimodal pattern emerges (Figure S1A). Most genes peak either in the early morning hours with a maximum around 05:00 (just before lights on) or in the late evening with a maximum around 20:00 (at lights off). This approach offers a simple determination of the peak times, but it is sensitive to noise, since a single outlier measurement can interfere with the determination of peak times. The second approach to determine the time of peak expression is based on correlating a shifted cosine curve of periodicity T = 24 h with the observed expression pattern. The peak time is identified as the time shift that maximizes the correlation. By this approach, we utilized all measurement points of the time series equally for the determination of the peak time and, thus, reduced the influence of outlier measurements. Furthermore, a higher temporal resolution could be achieved (Figure S1B). Although differences were observed for some genes, the resulting distributions of peak times were similar for both approaches (Figure S1). This indicates that the influence of outlier measurements was minor in our experiment and points to a general high quality of data. The differences between both approaches can also be seen by visualizing the ordered expression matrices (data not shown). It appears that the second approach leads to a 'smoother' ordering of the temporal expression profiles and, thus, may be favorable in cases in which genes should be sorted according to their transcription patterns.

To obtain an estimate of the number of expressed genes measured by the microarrays, we utilized the arrays' unique feature that they included probes for *Prochlorococcus* MIT9313 and several phages besides probes for *Prochlorococcus* MED4. As we did not expect to measure expression for most of phages genes in the experiment, the corresponding phage probes sets were used for an estimation of the background intensity for non-expressed genes. First, the median signal intensity was calculated for each probe set. A crude threshold for expression was subsequently defined by determining the 0.95-quantile for signals of phages probe sets i.e. the threshold for which 95% of the phage signal intensities lie below. This threshold was chosen as we expected (and observed) that a small percentage of phage probe set will still display large expression values due to cross-hybridization with homologous genes or hybridization artifacts. The threshold obtained (29.9 arbitrary units) was then used to classify MED4 probes as "expressed" or "non-expressed." For the following analyses, genes were included if they met one of two criteria: significant periodicity over the diel cycle (FDR<0.10), or, for the aperiodic genes (FDR≥0.10), being classified as expressed.

To examine the relative temporal expression patterns for the periodic genes, soft clustering was applied. In contrast to conventional (hard) clustering such as k-means (where genes belong to exactly one cluster), the memberships of genes to clusters were graded between 0 and 1. Large membership values imply that the genes were strongly associated with the cluster; low membership implies that the genes were poorly represented by the cluster. Soft clustering offers the advantage of producing information-rich clustering structure and of being more robust to noise [112]. For the cluster analysis, the Bioconductor package *Mfuzz* was used [113]. The clustering parameter m determining the 'softness' of the cluster was set to 1.25. The appropriate cluster number c was difficult to determine for this data set since there are two dominant expression patterns (corresponding to the genes peaking in the morning or evening, respectively). These two major clusters can, however, be further subdivided. Successive clustering with increasing cluster number reflected this finding showing first the main expression patterns and subsequently the minor patterns. To obtain an optimal cluster number, we assessed the functional enrichment of detected clusters varying the cluster number [30]. Consequently, the cluster number was set to 16, as it maximized the total number of subcategories of functional genes (see below) enriched for the transcriptome (data not shown).

To interpret the biological significance of the observed expression patterns, we examined the clusters obtained for enrichment of genes with known function. For this task, we utilised the functional categorization of *Prochlorococcus marinus* MED4 by the Cyanobase (*http://www.kazusa.or.jp/cyano/*) where 1193 genes are associated with 16 main and 62 sub-categories [114]. Of the 1193 annotated genes, 820 were found expressed in the experiment. Subsequently, we used this set of genes to associate possible functions to the expression patterns observed. The statistical significance of observing k genes of a defined function in a cluster with a total of l genes can be derived from the hyper-geometrical distribution

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{l-i}}{\binom{N}{M}}$$

where M is the total number of genes attributed to the function of interest, N is the total number of genes annotated and P is the probability to observe k or more genes of the function of interest if they would be randomly drawn. Since multiple testing was performed, the p-values obtained were adjusted using the Benjamini-Hochberg procedure [115].

Supporting Information

Table S1 Expression profiles of all MED4 probe sets. Open reading frames ("PMM####") intergenic regions

("PMMIG…") and non-coding RNAs ("PMM_…") are listed in column 1, followed by annotations in column 3. PMM#### in column 1 are the annotations that were deposited in Genbank when MED4 was first sequenced [3]. Subsequently, with the sequencing of more *Prochlorococcus* strains, the genes have been renamed [2], and this new nomenclature is shown in column 2. Fourier score (column 4) and false discovery rate (FDR) for the score (column 5) are followed by calculated peak expression time (column 6) and calculated Pearson correlation with a (possibly) shifted cosine curve (column 7). Cluster assignment (column 8) and cluster membership (column 9) are followed by Cyanobase functional category (column 10) and sub-category (column 11) assignments. The final 100 columns list the mean RMA-normalized expression and the standard deviation of the mean of the 50 time points.
Found at: doi:10.1371/journal.pone.0005135.s001 (4.22 MB XLS)

Table S2
Found at: doi:10.1371/journal.pone.0005135.s002 (0.07 MB DOC)

Table S3
Found at: doi:10.1371/journal.pone.0005135.s003 (0.10 MB DOC)

Table S4
Found at: doi:10.1371/journal.pone.0005135.s004 (0.24 MB DOC)

Table S5
Found at: doi:10.1371/journal.pone.0005135.s005 (0.09 MB DOC)

Table S6
Found at: doi:10.1371/journal.pone.0005135.s006 (0.11 MB DOC)

Table S7
Found at: doi:10.1371/journal.pone.0005135.s007 (0.07 MB DOC)

Table S8
Found at: doi:10.1371/journal.pone.0005135.s008 (0.10 MB DOC)

Table S9
Found at: doi:10.1371/journal.pone.0005135.s009 (0.16 MB DOC)

Table S10
Found at: doi:10.1371/journal.pone.0005135.s010 (0.12 MB DOC)

Table S11
Found at: doi:10.1371/journal.pone.0005135.s011 (0.06 MB DOC)

Figure S1 Histograms of the peak expression time of all periodic genes by (A) averaging over both days or (B) correlating a shifted cosine curve of periodicity T = 24 h. See text for details. Relative PAR (photosynthetically available radiation) over the experiment is represented above the histograms.
Found at: doi:10.1371/journal.pone.0005135.s012 (1.36 MB EPS)

Figure S2 Members of the 16 periodic clusters (Clusters 1–16) as well as the aperiodic expressed (Cluster 17) and the unexpressed (Cluster 18) clusters. Periodic clusters were assigned and arranged chronologically, with the mean peak abundance of Cluster 1 (08:20) occurring the soonest after experimental dawn (06:00), and Cluster 16 occurring the latest (05:30). Red color indicates strong group membership (i.e high fuzziness score), while yellow indicates weak group membership. Relative PAR (photosynthetically available radiation) over the experiment is represented above the expression patterns.
Found at: doi:10.1371/journal.pone.0005135.s013 (9.07 MB EPS)

Figure S3 Expression profiles of representative genes of MED4 during a pilot 12 hour light - 12 hour dark experiment, monitored by quantitative reverse transcription PCR. Values are expressed as the ratio of gene expression versus expression of the aperiodic gene *rnpB*. For all genes, including *rnpB*, and all time points, the coefficient of variation for replicate PCR reactions was less than 7.0.
Found at: doi:10.1371/journal.pone.0005135.s014 (2.48 MB EPS)

Figure S4 Distribution of Fourier scores. To assess the significance of periodic expression, the distribution of Fourier scores (red line) for the measured gene expression was compared with the distribution obtained for AR(1)-based background models (blue line). The false discovery rate (orange line) denotes the fraction of Fourier scores derived from background distribution in respect to the number of scores in the observed distribution above a chosen threshold.
Found at: doi:10.1371/journal.pone.0005135.s015 (0.94 MB EPS)

Acknowledgments

We thank George Church, Kyriacos Leptos, Xiaoxia Lin, and other members of the Church laboratory, and Julia Holtzendorff and Gabrielle Rocap for advice and technical assistance.

Author Contributions

Conceived and designed the experiments: ERZ DL ZJ SWC. Performed the experiments: ERZ DL ZJ CS MC NM. Analyzed the data: ERZ DL ZJ MEF CS MC LRT SWC. Contributed reagents/materials/analysis tools: MEF MC MAW TR RS SWC. Wrote the paper: ERZ DL ZJ MEF CS MC LRT SWC.

References

1. Partensky F, Hess WR, Vaulot D (1999) *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. Microbiol Molec Biol Rev 63: 106–127.
2. Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, et al. (2007) Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. PLoS Genetics 3: 2515–2528.
3. Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, et al. (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature 424: 1042–1047.
4. Steglich C, Futschik ME, Lindell D, Voss B, Chisholm SW (2008) The challenge of regulation in a minimal photoautotroph: Non-coding RNAs in *Prochlorococcus*. PLoS Genetics 4: e1000173.
5. Bruyant F, Babin M, Genty B, Prasil O, Behrenfeld MJ, et al. (2005) Diel variations in the photosynthetic parameters of *Prochlorococcus* strain PCC 9511: Combined effects of light and cell cycle. Limnol Oceanogr 50: 850–863.
6. Claustre H, Bricaud A, Babin M, Bruyant F, Guillou L, et al. (2002) Diel variations in *Prochlorococcus* optical properties. Limnol Oceanogr 47: 1637–1647.
7. Garczarek L, Partensky F, Irlbacher H, Holtzendorff J, Babin M, et al. (2001) Differential expression of antenna and core genes in *Prochlorococcus* PCC 9511 (Oxyphotobacteria) grown under a modulated light-dark cycle. Environ Microbiol 3: 168–175.
8. Pichard SL, Campbell L, Kang JB, Tabita FR, Paul JH (1996) Regulation of ribulose bisphosphate carboxylase gene expression in natural phytoplankton communities. I. Diel rythms. Mar Ecol Prog Ser 139: 257–265.
9. Liu HB, Nolla HA, Campbell L (1997) *Prochlorococcus* growth rate and contribution to primary production in the equatorial and subtropical North Pacific Ocean. Aquat Microb Ecol 12: 39–47.
10. Partensky F, Blanchot J, Lantoine F, Neveux J, Marie D (1996) Vertical structure of picophytoplankton at different trophic site of the subtropical northeastern Atlantic. Deep-Sea Res I 43: 1191–1213.

11. Vaulot D, Marie D, Olson RJ, Chisholm SW (1995) Growth of *Prochlorococcus*, a photosynthetic prokaryote, in the equatorial Pacific Ocean. Science 268: 1480–1482.
12. Holtzendorff J, Partensky F, Mella D, Lennon JF, Hess WR, et al. (2008) Genome streamlining results in loss of robustness of the circadian clock in the marine cyanobacterium *Prochlorococcus marinus* PCC 9511. J Biol Rhythms 23: 187–199.
13. Shalapyonok A, Olson RJ, Shalapyonok LS (1998) Ultradian growth in *Prochlorococcus* spp. Appl Environ Microbiol 64: 1066–1069.
14. Burbage CD, Binder BJ (2007) Relationship between cell cycle and light-limited growth rate in oceanic *Prochlorococcus* (MIT9312) and *Synechococcus* (WH8103) (cyanobacteria). J Phycol 43: 266–274.
15. Jacquet S, Partensky F, Lennon J-F, Vaulot D (2001) Diel patterns of growth and division in marine picoplankton in culture. J Phycol 37: 357–369.
16. Holtzendorff J, Partensky F, Jacquet S, Bruyant F, Marie D, et al. (2001) Diel expression of cell cycle-related genes in synchronized cultures of *Prochlorococcus* PCC 9511. J Bacteriol 183: 915–920.
17. Chen YB, Dominic B, Mellon MT, Zehr JP (1998) Circadian rhythm of nitrogenase gene expression in the diazotrophic filamentous nonheterocystis cyanobacterium *Trichodesmium* sp. strain IMS 101. J Bacteriol 180: 3598–3605.
18. Golden SS (2003) Timekeeping in bacteria: The cyanobacterial circadian clock. Curr Opin Microbiol 6: 535–540.
19. Liu Y, Tsinoremas N, Hirschie Johnson C, Lebedeva N, Golden SS, et al. (1995) Circadian orchestration of gene expression in cyanobacteria. Genes Dev 9: 1469–1478.
20. Sweeney BM, Borgese MB (1989) A circadian rhythm in cell division in a prokaryote, the cyanobacterium *Synechococcus* WH7803. J Phycol 25: 183–186.
21. Woelfle MA, Johnson CH (2006) No promoter left behind: Global circadian gene expression in cyanobacteria. J Biol Rhythms 21: 419–431.
22. Golden SS, Canales SR (2003) Cyanobacterial circadian clocks - Timing is everything. Nature Rev Microbiol 1: 191–199.
23. Takai N, Nakajima M, Oyama T, Kito R, Sugita C, et al. (2006) A KaiC-associating SasA-RpaA two-component regulatory system as a major circadian timing mediator in cyanobacteria. Proc Natl Acad Sci USA 103: 12109–12114.
24. Smith RM, Williams SB (2006) Circadian rhythms in gene transcription imparted by chromosome compaction in the cyanobacterium *Synechococcus elongatus*. Proc Natl Acad Sci USA 103: 8564–8569.
25. Ivleva NB, Gao T, LiWang AC, Golden SS (2006) Quinone sensing by the circadian input kinase of the cyanobacterial circadian clock. Proc Natl Acad Sci USA 103: 17468–17473.
26. West NJ, Scanlan DJ (1999) Niche-partitioning of *Prochlorococcus* populations in a stratified water column in the eastern North Atlantic Ocean. Appl Environ Microbiol 65: 2585–2591.
27. Lindell D, Jaffe JD, Coleman ML, Futschik ME, Axmann IM, et al. (2007) Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. Nature 449: 83–86.
28. Martiny AC, Coleman ML, Chisholm SW (2006) Phosphate acquisition genes in *Prochorococcus* ecotypes: Evidence for genome-wide adaptation. Proc Natl Acad Sci USA 103: 12552–12557.
29. Steglich C, Futschik M, Rector T, Steen R, Chisholm S (2006) Genome-wide analysis of light sensing in *Prochlorococcus*. J Bacteriol 188: 7796–7806.
30. Tolonen AC, Aach J, Lindell D, Johnson ZI, Rector T, et al. (2006) Global gene expression of *Prochlorococcus* ecotypes in response to changes in nitrogen availability. Mol Syst Biol 2: Article 53.
31. Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, et al. (2006) Genomic islands and the ecology and evolution of *Prochlorococcus*. Science 311: 1768–1770.
32. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, et al. (2007) The *Sorcerer II* Global ocean sampling expedition: Northwest Atlantic through eastern tropical Pacific. PLoS Biology 5: 398–431.
33. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304: 66–74.
34. Dufresne A, Garczarek L, Partensky F (2005) Accelerated evolution associated with genome reduction in a free-living prokaryote. Genome Biol 6: R14.
35. Vicente M, Rico AI, Martinez-Arteaga R, Mingorance J (2006) Septum enlightenment: Assembly of bacterial division proteins. J Bacteriol 188: 19–27.
36. Rothfield L, Taghbalout A, Shih YL (2005) Spatial control of bacterial division-site placement. Nature Rev Microbiol 3: 959–968.
37. Kornberg A, Baker T (1992) DNA Replication. New York: W.H. Freeman and Co.
38. Campbell L, Nolla HA, Vaulot D (1994) The importance of *Prochlorococcus* to community structure in the central North Pacific Ocean. Limnol Oceanogr 39: 954–961.
39. Talling JF (1966) Photosynthetic behavior in stratified and unstratified lake populations of a planktonic diatom. J Ecol 54: 99–127.
40. Behrenfeld MJ, Prasil O, Babin M, Bruyant F (2004) In search of a physiological basis for covariations in light-limited and light-saturated photosynthesis. J Phycol 40: 4–25.
41. Kucho K, Okamoto K, Tsuchiya Y, Nomura S, Nango M, et al. (2005) Global analysis of circadian expression in the cyanobacterium *Synechocystis* sp. strain PCC 6803. J Bacteriol 187: 2190–2199.
42. Labiosa RG, Arrigo KR, Tu CJ, Bhaya D, Bay S, et al. (2006) Examination of diel changes in global transcript accumulation in *Synechocystis* (cyanobacteria). J Phycol 42: 622–636.
43. Stockel J, Welsh EA, Liberton M, Kunnvakkam R, Aurora R, et al. (2008) Global transcriptomic analysis of *Cyanothece* 51142 reveals robust diurnal oscillation of central metabolic processes. Proc Nat Acad Sci USA 105: 6156–6161.
44. Toepel J, Welsh E, Summerfield TC, Pakrasi HB, Sherman LA (2008) Differential transcriptional analysis of the cyanobacterium *Cyanothece* sp. strain ATCC 51142 during light-dark and continuous light growth. J Bacteriol 190: 3904–3913.
45. Aro E-M, Virgin I, Andersson B (1993) Photoinhibition of photosystem II. Inactivation, protein damage, and turnover. Biochim Biophys Acta 1143: 113–134.
46. Baroli I, Melis A (1998) Photoinhibitory damage is modulated by the rate of photosynthesis and by the photosystem II light-harvesting chlorophyll antenna size. Planta 205: 288–296.
47. Osmond CB (1994) What is photoinhibition? Some insights from comparisons of shade and sun plants. In: Baker NR, Bowyer JR, eds (1994) Photoinhibition of Photosynthesis: from molecular mechanisms to the field. Oxford, UK: BIOS Scientific Publishers Limited. pp 1–24.
48. Diner BA, Rappaport F (2002) Structure, dynamics, and energetics of the primary photochemistry of photosystem II of oxygenic photosynthesis. Ann Rev Plant Biol 53: 551–580.
49. Bustos SA, Schaefer MR, Golden SS (1990) Differential and rapid responses of four cyanobacterial *psbA* transcripts to changes in light intensity. J Bacteriol 172: 1998–2004.
50. Schaefer MR, Golden SS (1989) Light availability influences the ratio of the two forms of D1 in cyanobacterial thylakoids. J Biol Chem 264: 7412–7417.
51. Behrenfeld MJ, Prasil O, Kolber ZS, Babin M, Falkowski PG (1998) Compensatory changes in Photosystem II electron turnover rates protect photosynthesis from photoinhibition. Photosynth Res 58: 259–268.
52. Garczarek L, van der Staay GWM, Hess WR, Le Gall F, Partensky F (2001) Expression and phylogeny of the multiple antenna genes of the low-light-adapted strain *Prochlorococcus marinus* SS120 (Oxyphotobacteria). Plant Mol Biol 46: 683–693.
53. Promnares K, Komenda J, Bumba L, Nebesarova J, Vacha F, et al. (2006) Cyanobacterial small chlorophyll-binding protein ScpD (HliB) is located on the periphery of photosystem II in the vicinity of PsbH and CP47 subunits. J Biol Chem 281: 32705–32713.
54. Wang Q, Jantaro S, Lu B, Majeed W, Bailey M, et al. (2008) The high light-inducible polypeptides stabilize trimeric photosystem I complex under high light conditions in *Synechocystis* PCC 6803. Plant Physiol 147: 1239–1250.
55. He Q, Dolganov N, Bjorkman O, Grossman AR (2001) The high light-inducible polypeptides in *Synechocystis* PCC6803. Expression and function in high light. J Biol Chem 276: 306–314.
56. Salem K, van Waasbergen LG (2004) Light control of *hliA* transcription and transcript stability in the cyanobacterium *Synechococcus elongatus* strain PCC 7942. J Bacteriol 186: 1729–1736.
57. Havaux M, Guedeney G, He Q, Grossman AR (2003) Elimination of high-light-inducible polypeptides related to eukaryotic chlorophyll *a*/*b*-binding proteins results in aberrant photoacclimation in *Synechocystis* PCC6803. Biochim Biophys Acta 1557: 21–33.
58. Xu H, Vavlin D, Funk C, Vermaas W (2004) Multiple deletions of small Cab-like proteins in the cyanobacterium *Synechocystis* sp. PCC 6803: consequences for pigment biosynthesis and accumulation. J Biol Chem 279: 27971–27979.
59. Bhaya D, Dufresne A, Vaulot D, Grossman A (2002) Analysis of the *hli* gene family in marine and freshwater cyanobacteria. FEMS Microbiol Lett 215: 209–219.
60. Lindell D, Sullivan MB, Johnson ZI, Tolonen AC, Rohwer F, et al. (2004) Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. Proc Nat Acad Sci USA 101: 11013–11018.
61. So AK-C, Espie GS, Williams EB, Shively JM, Heinhorst S, et al. (2004) A Novel Evolutionary Lineage of Carbonic Anhydrase (ε Class) Is a Component of the Carboxysome Shell. J Bacteriol 186: 623–630.
62. Howitt CA, Udall PK, Vermaas WFJ (1999) Type 2 NADH dehydrogenases in the cyanobacterium *Syenchocystis* sp. Strain PCC 6803 are involved in regulation rather than respiration. J Bacteriol 181: 3994–4003.
63. Badger MR, Price GD, Long BM, Woodger FJ (2006) The environmental plasticity and ecological genomics of the cyanobacterial CO_2 concentrating mechanism. J Exper Bot 57: 249–265.
64. Battchikova N, Aro EM (2007) Cyanobacterial NDH-1 complexes: multiplicity in function and subunit composition. Physiol Plant 131: 22–32.
65. Bukhov N, Carpentier R (2004) Alternative photosystem I-driven electron transport routes: mechanisms and functions. Photosynth Res 82: 17–33.
66. Goldberg RN, Tewari YB, Bhat TN (2004) Thermodynamics of enzyme-catalyzed reactions–a database for quantitative biochemistry. Bioinformatics 20: 2874–2877.
67. Tamoi M, Miyazaki T, Fukamizo T, Shigeoka S (2005) The Calvin cycle in cyanobacteria is regulated by CP12 via the NAD(H)/NADP(H) ratio under light/dark conditions. Plant J 42: 504–513.
68. Wedel N, Soll J, Paap BK (1997) CP12 provides a new mode of light regulation of Calvin cycle activity in higher plants. Proc Nat Acad Sci USA 94: 10479–10484.

69. Hagen KD, Meeks JC (2001) The unique cyanobacterial protein OpcA is an allosteric effector of glucose-6-phosphate dehydrogenase in *Nostoc punctiforme* ATCC 29133. J Biol Chem 276: 11477–11486.
70. Bertilsson S, Berglund O, Karl DM, Chisholm SW (2003) Elemental composition of marine *Prochlorococcus* and *Synechococcus*: Implications for the ecological stoichiometry of the sea. Limnol Oceanogr 48: 1721–1731.
71. Van Mooy BAS, Rocap G, Fredricks HF, Evans CT, Devol AH (2006) Sulfolipids dramatically decrease phosphorus demand by picocyanobacteria in oligotrophic marine environments. Proc Natl Acad Sci USA 103: 8607–8612.
72. Moore LR, Ostrowski M, Scanlan DJ, Feren K, Sweetsir T (2005) Ecotypic variation in phosphorus-acquisition mechanisms within marine picocyanobacteria. Aquat Microb Ecol 39: 257–269.
73. Muro-Pastor MI, Reyes JC, Florencio FJ (2005) Ammonium assimilation in cyanobacteria. Photosynth Res 83: 135–150.
74. Moore LR, Post AF, Rocap G, Chisholm SW (2002) Utilization of different nitrogen sources by the marine cyanobacteria, *Prochlorococcus* and *Synechococcus*. Limnol Oceanogr 47: 989–996.
75. Mary I, Garczarek L, Tarran GA, Kolowrat C, Terry MJ, et al. (2008) Diel rhythmicity in amino acid uptake by *Prochlorococcus*. Environ Microbiol 10: 2124–2131.
76. Iwasaki H, Williams SB, Kitayama Y, Ishiura M, Golden SS, et al. (2000) A KaiC-interacting sensory histidine kinase, SasA, necessary to sustain robust circadian oscillation in cyanobacteria. Cell 101: 223–233.
77. Katayama M, Tsinoremas NF, Kondo T, Golden SS (1999) *cpmA*, a gene involved in an output pathway of the cyanobacterial circadian system. J Bacteriol 181: 3516–3524.
78. Nair U, Ditty JL, Min H, Golden SS (2002) Roles for sigma factors in global circadian regulation of the cyanobacterial genome. J Bacteriol 184: 3530–3538.
79. Tsinoremas N, Ishiura M, Kondo T, Andersson C, Tanaka K, et al. (1996) A sigma factor that modifies the circadian expression of a subset of genes in cyanobacteria. The EMBO Journal 15: 2488–2495.
80. Imamura S, Asayama M, Takahashi H, Tanaka K, Takahashi H, et al. (2003) Antagonistic dark/light-induced SigB/SigD, group 2 sigma factors, expression through redox potential and their roles in cyanobacteria. FEBS Lett 554: 357–362.
81. Yoshimura T, Imamura S, Tanaka K, Shirai M, Asayama M (2007) Cooperation of group 2 σ factors, SigD and SigE for light-induced transcription in the cyanobacterium *Synechocystis* sp. PCC 6803. FEBS Lett 581: 1495–1500.
82. Summerfield TC, Sherman LA (2007) Role of sigma factors in controlling global gene expression in light/dark transitions in the cyanobacterium *Synechocystis* sp. strain PCC 6803. J Bacteriol 189: 7829–7840.
83. Figge RM, Cassier-Chauvat C, Chauvat F, Cerff R (2000) Characterization and analysis of an NAD(P)H dehydrogenase transcriptional regulator for the survival of cyanobacteria facing inorganic carbon starvation and osmotic stress. Mol Microbiol 39: 455–468.
84. Espinosa J, Forchhammer K, Burillo S, Contreras A (2006) Interaction network in cyanobacterial nitrogen regulation: PipX, a protein that interacts in a 2-oxoglutarate dependent manner with PII and NtcA. Mol Microbiol 61: 457–469.
85. Moore LR, Coe A, Zinser ER, Saito MA, Sullivan MB, et al. (2007) Culturing the marine cyanobacterium *Prochlorococcus*. Limnol Oceanogr: Methods 5: 353–362.
86. Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, et al. (2008) Microbial community gene expression in ocean surface waters. Proc Nat Acad Sci USA 105: 3805–3810.
87. Li H, Sherman LA (2000) A redox-responsive regulator of photosynthesis gene expression in the cyanobacterium *Synechocystis* sp. strain PCC 6803. J Bacteriol 182: 4268–4277.
88. Liu Y, Golden SS, Kondo T, Ishiura M, Hirschie Johnson C (1995) Bacterial luciferase as a reporter of circadian gene expression in cyanobacteria. J Bacteriol 177: 2080–2086.
89. Ouyang Y, Andersson CR, Kondo T, Golden SS, Johnson CH (1998) Resonating circadian clocks enhance fitness in cyanobacteria. Proc Natl Acad Sci USA 95: 8660–8664.
90. Follows MJ, Dutkiewicz S, Grant S, Chisholm SW (2007) Emergent biogeography of microbial communities in a model ocean. Science 315: 1843–1846.
91. Lindell D, Jaffe JD, Johnson ZI, Church GM, Chisholm SW (2005) Photosynthesis genes in marine viruses yield proteins during host infection. Nature 438: 86–89.
92. Vaulot D, Courties C, Partensky F (1989) A simple method to preserve oceanic phytoplankton for flow cytometric analyses. Cytometry 10: 629–635.
93. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, et al. (2002) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4: 249–264.
94. Binder BJ, Chisholm SW, Olson RJ, Frankel SL, Worden AZ (1996) Dynamics of picophytoplankton, ultraphytoplankton, and bacteria in the central equatorial Pacific. Deep-Sea Res II 43: 907–931.
95. Monger BC, Landry MR (1993) Flow cytometric analysis of marine bacteria with Hoechst 33342. Appl Environ Microbiol 59: 905–911.
96. Cavender-Bares KK, Mann EL, Chisholm SW, Ondrusek ME, Bidigare RR (1999) Differential response of equatorial Pacific phytoplankton to iron fertilization. Limnol Oceanogr 44: 237–246.
97. Olson RJ, Vaulot D, Chisholm SW (1985) Marine-Phytoplankton Distributions Measured Using Shipboard Flow-Cytometry. Deep-Sea Res I 32: 1273–1280.
98. Carpenter EJ, Chang J (1988) Species-specific phtyoplankton growth rates via diel DNA synthesis cycles. I. Concept of the method. Mar Ecol Prog Ser 43: 105–111.
99. Lewis MR, Smith K (1983) A small volume, short-incubation-time method for measurement of photosynthesis as a function of incident irradiance. Mar Ecol Prog Ser 13: 99–102.
100. Johnson Z, Landry ML, Bidigare RR, Brown SL, Campbell L, et al. (1999) Energetics and growth kinetics of a deep *Prochlorococcus* spp. population in the Arabian Sea. Deep-Sea Res II 46: 1719–1743.
101. Johnson ZI, Sheldon TL (2007) A high-throughput method to measure photosynthesis-irradiance curves of phytoplankton. Limnol Oceanogr: Methods 5: 417–424.
102. Barber RT, Sanderson MP, Lindley ST, Chai F, Newton J, et al. (1996) Primary productivity and its regulation in the equatorial Pacific during and following the 1991–1992 El Nino. Deep-Sea Res II 43: 933–969.
103. Webb WL, Newton M, Star D (1974) Carbon dioxide exchange of *Alnus rubra*: a mathematical model. Oecologica 17: 281–291.
104. Johnson Z, Barber R (2003) The low-light reduction in the quantum yield of photosynthesis: potential errors and biases when calculating the maximum quantum yield. Photosynth Res 75: 85–95.
105. Sakshaug E, Bricaud A, Dandonneau Y, Falkowski PG, Kiefer DA, et al. (1997) Parameters of photosynthesis: definitions, theory and interpretation of results. J Plankton Res 19: 1637–1670.
106. Johnson ZI (2004) Development and application of the Background Irradiance Gradient - Single Turnover Fluorometer (BIG-STf). Mar Ecol Prog Ser 283: 73–80.
107. Kolber ZS, Prasil O, Falkowski PG (1998) Measurements of Variable Chlorophyll Fluorescence Using Fast Repetition Rate Techniques - Defining Methodology and Experimental Protocols. Biochim Biophys Acta 1367: 88–106.
108. Irizarry RA, Bolstad B, Collin L, Cope L, Hobbs B, et al. (2003) Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 31: e15.
109. Choe SE, Boutros M, Michelson AM, Church GM, Halfon MS (2005) Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. Genome Biol 6: R2.
110. de Lichtenberg U, Jensen LJ, Fausboll A, Jensen TS, Bork P, et al. (2005) Comparison of computational methods for the identification of cell cycle-regulated genes. Bioinformatics 21: 1164–1171.
111. Futschik ME, Herzel H (2008) Are we overestimating the number of cell-cycling genes? The impact of background models on time series analysis. Bioinformatics 23: 605–611.
112. Futschik ME, Charlisle B (2005) Noise robust clustering of gene expression time-course data. J Bioinform Comp Biol 3: 965–988.
113. Kumar L, Futschik ME (2007) Mfuzz: a software package for soft clustering of microarray data. Bioinformation 2: 5–7.
114. Nakamura Y, Kaneko T, Hirosawa M, Miyajima N, Tabata S (1998) Cyanobase, a www database containing the complete nucleotide sequence of the genome of *Synechocystis* sp. strain PCC6803. Nucleic Acids Res 26: 63–67.
115. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate - A practical and powerful approach to multiple testing. J R Statist Soc B 57: 289–300.

Table S1: Expression profiles of all MED4 probe sets.
(first page only – see full table at http://www.plosone.org/article/info:doi/10.1371/journal.pone.0005135)

											0	0	2	2	4	4
Gene or region	New gene name	Function	Fourier	FDR	Peak	Peak R value	Cluster	Cluster membership score	Cyanobase categories, main	Cyanobase sub-categories	Normalized	StdDev Norm	Normalized	StdDev Norm	Normalized	StdDev Norm
PMM0001	PMED4_00001	dnaN DNA polymerase III, beta chain	1.605377599	0.000470588	18.2	0.917160314	5	0.672061576	11.0 DNA replication, recombination, and repair	11.0 DNA replication, recombination, and repair	0.38633642	0.008823202	0.3674509	0.11788837	0.96665746	0.29353213
PMM0002	PMED4_00011	conserved hypothetical protein	1.496153842	0.002642706	18	0.855513467	6	0.607841656	16.0 Hypothetical	16.1 Conserved hypothetical protein	1.124593	0.16195475	0.48498905	0.06167883	0.89124227	0.001094597
PMM0003	PMED4_00021	purL phosphoribosylformylglycinamidine synthetase II	1.633184477	0.000325733	18.4	0.931530392	6	0.997901284	9.0 Purines, pyrimidines, nucleosides, and nucleotides	9.2 Purine ribonucleotide biosynthesis	0.83767873	0.019058136	0.6134913	0.058455907	0.89524627	0.009508674
PMM0004	PMED4_00031	purF Glutamine amidotransferase class-II	1.496142265	0.002639916	21.8	0.854838364	9	0.76781931	9.0 Purines, pyrimidines, nucleosides, and nucleotides	9.2 Purine ribonucleotide biosynthesis	1.035698	0.17598955	0.68168706	0.020902	0.81449893	0.013220658
PMM0005	PMED4_00041	DNA gyrase/topoisomerase IV, subunit A	1.593096264	0.000421053	18	0.911298025	6	0.631469693	11.0 DNA replication, recombination, and repair	11.0 DNA replication, recombination, and repair	0.6844061	0.034790408	0.6090897	0.018041862	0.891863	0.03978797
PMM0006	PMED4_00051	conserved hypothetical protein	1.522402536	0.000952381	18.8	0.866181825	6	0.972469084	16.0 Hypothetical	16.1 Conserved hypothetical protein	0.7769074	0.06906239	0.42945316	0.02118187	0.83312243	0.010860079
PMM0007	PMED4_00061	conserved hypothetical protein	1.362970706	0.015112994	19.6	0.77312412	8	0.439104915	16.0 Hypothetical	16.1 Conserved hypothetical protein	0.82526636	0.18957119	0.5401521	0.094155066	0.6319684	0.10673177
PMM0008	PMED4_00071	conserved hypothetical protein	1.544879357	0.000824176	18.8	0.879042759	6	0.949065779	16.0 Hypothetical	16.1 Conserved hypothetical protein	0.9062444	0.068989664	0.6384533	0.11340126	0.86576587	0.14318147
PMM0009	PMED4_00081	nusB Antitermination protein NusB	1.499873566	0.002462527	20.2	0.85038616	7	0.694453624	12.0 Transcription	12.2 RNA synthesis, modification, and DNA transcription	0.9709103	0.08896643	0.6099937	0.026181588	0.69248617	0.055825453
PMM0010	PMED4_00091	ftsY signal recognition particle docking protein FtsY	1.454980661	0.005369718	19.4	0.825717776	6	0.654253974	4.0 Cellular processes	4.1 Cell division	0.90520316	0.08246207	0.7609995	0.01892876	0.82533634	0.12555487
PMM0011	PMED4_00101	Protein phosphatase 2C domain	9.214921202	0.229278351	17.4	0.530066014	18	1	10.0 Regulatory functions	10.0 Regulatory functions	1.1539766	0.008745134	0.75928575	0.12382859	0.8103816	0.036645334
PMM0012	PMED4_00111	argH Fumarate lyase	1.661888367	0	18.2	0.949170029	6	0.518133679	1.0 Amino acid biosynthesis	1.4 Glutamate family / Nitrogen assimilation	0.70565987	0.057005934	0.5213482	0.014700573	0.93627995	0.049191583
PMM0013	PMED4_00121	RNA-binding region RNP-1 (RNA recognition motif)	1.408822186	0.008752905	4.8	0.814535155	16	0.790414156	7.0 Fatty acid, phospholipid and sterol metabolism	7.0 Fatty acid, phospholipid and sterol metabolism	1.113227	0.055423293	1.0981033	0.024064062	1.0646299	0.15553854
PMM0014	PMED4_00131	conserved hypothetical protein	13.8730048	0.011603843	23	0.800274843	9	0.908516299	16.0 Hypothetical	16.1 Conserved hypothetical protein	1.177626	0.08247345	0.7120032	0.037706908	0.7946318	0.037393954
PMM0015	PMED4_00141	Domain of unknown function DUF25	10.2828567	0.148897886	12.2	0.600813299	17	1	16.0 Hypothetical	16.1 Conserved hypothetical protein	1.1573404	0.12991348	0.975642	0.023749987	1.0380604	0.18995507
PMM0016	PMED4_00151	grpE Heat shock protein GrpE	1.252147249	0.039264877	22	0.71585645	8	0.86939086	4.0 Cellular processes	4.3 Chaperones	0.6227597	0.17204303	0.3643536	0.008778083	0.35401827	0.038542304
PMM0017	PMED4_00161	dnaJ Dnaj protein	1.493129508	0.003215768	21.4	0.850030004	8	0.980173266	4.0 Cellular processes	4.3 Chaperones	0.7745841	0.034263536	0.49795872	0.0877481	0.6200356	0.02904216
PMM0018	PMED4_00171	conserved hypothetical protein	1.302984543	0.025094817	22.6	0.749333369	9	0.961940403	16.0 Hypothetical	16.1 Conserved hypothetical protein	1.242183	0.17869323	0.71350986	0.21729001	0.64111006	0.11780182
PMM0019	PMED4_00181	conserved hypothetical protein	1.404949462	0.009056025	20.4	0.796861648	8	0.465676101	16.0 Hypothetical	16.1 Conserved hypothetical protein	1.0200331	0.027423028	0.6054563	0.11844452	0.8627308	0.03032723
PMM0020	PMED4_00191	conserved hypothetical protein	1.604196325	0.000465116	18	0.917423867	5	0.979853057	16.0 Hypothetical	16.1 Conserved hypothetical protein	0.5268917	0.008982025	0.55634654	0.10538431	1.0543964	0.13290393
PMM0021	PMED4_00201	murB UDP-N-acetylenolpyruvoylglucosamine reductase	1.119025901	0.093922925	19.6	0.634595755	6	0.355074601	3.0 Cell envelope	3.2 Murein sacculus and peptidoglycan	0.90302855	0.07635757	0.8333424	0.062261753	0.8370844	0.036232725
PMM0022	PMED4_00211	murC Probable UDP-N-acetylmuramate-alanine ligase	1.464945061	0.004562044	17.6	0.840909623	5	0.681563646	3.0 Cell envelope	3.2 Murein sacculus and peptidoglycan	0.9896275	0.004605908	0.6753201	0.07521578	0.8482649	0.122346714
PMM0023	PMED4_00221	gap2 Glyceraldehyde 3-phosphatedehydrogenase (NADP+)(phosphorylating)	1.622383514	0.000547945	6.8	0.923086475	16	0.564184029	8.0 Photosynthesis and respiration	8.2 CO2 fixation	2.2118858	0.014848654	1.7231948	0.08930658	1.4573263	0.25655392
PMM0024	PMED4_00231	thiL putative thiamine-monophosphate kinase	1.125420982	0.09112662	5	0.649908002	16	0.39254174	2.0 Biosynthesis of cofactors, prosthetic groups, and carriers	2.11 Thiamine	1.2040733	0.001596783	1.1429468	0.2290523	1.071014	0.038269803
PMM0025	PMED4_00241	Cyclophilin-type peptidyl-prolyl cis-transisomerase	1.665278614	0	6.2	0.95035777	16	0.946557318	13.0 Translation	13.5 Protein modification and translation factors	1.3817904	0.101495445	1.2438135	0.17447229	1.0243917	0.16881359
PMM0026	PMED4_00251	efp Elongation factor P (EF-P)	1.398157137	0.010355253	8	0.792551136	1	0.914684914	13.0 Translation	13.5 Protein modification and translation factors	1.3158221	0.046144307	1.1517549	0.021429026	1.2031878	0.15757194
PMM0027	PMED4_00261	accB,fabE Biotin / Lipoyl attachment	1.329707901	0.020159151	2	0.783434951	11	0.415734332	7.0 Fatty acid, phospholipid and sterol metabolism	7.0 Fatty acid, phospholipid and sterol metabolism	1.1040446	0.02635204	0.94763726	0.09932931	0.8678716	0.17539236
PMM0028	PMED4_00271	pdxA putative pyridoxal phosphate biosynthetic protein PdxA	1.049905049	0.135516605	21	0.59648219	17	1	2.0 Biosynthesis of cofactors, prosthetic groups, and carriers	2.9 Pyridoxine	0.9350901	0.099093035	0.606505	0.15332332	0.86560786	0.13411547
PMM0029	PMED4_00281	conserved hypothetical protein	1.584279415	0.958744016	4.6	0.091899923	17	1	16.0 Hypothetical	16.1 Conserved hypothetical protein	1.2747662	0.20628133	0.9477293	0.050932866	1.0101744	0.13466512
PMM0030	PMED4_00291	possible Transcription factor TFIID (or TATA-b	12.3538932	0.045902579	18.2	0.705828762	6	0.604322727	13.0 Translation	13.5 Protein modification and translation factors	0.67487967	0.09124684	0.44243008	0.01917504	0.84787476	0.13276573
PMM0031	PMED4_00301	HNH endonuclease	1.581500032	0.000369686	17.2	0.911652848	5	0.984961373	11.0 DNA replication, recombination, and repair	11.0 DNA replication, recombination, and repair	0.7812468	0.018567936	0.5805699	0.044397563	0.9958491	0.14379892
PMM0032	PMED4_00311	possible Bacterial type II secretion system pr	1.622905658	0.000274725	18.2	0.926245737	5	0.555074704	4.0 Cellular processes	4.6 Protein and peptide secretion	0.3686121	0.09093449	0.36849567	0.034676984	0.72505516	0.109908484
PMM0033	PMED4_00321	conserved hypothetical protein	16.3098605	0.00031348	18.6	0.928317526	6	0.944568415	16.0 Hypothetical	16.1 Conserved hypothetical protein	0.6554526	0.057313513	0.54595274	0.10168146	0.941091	0.051457528
PMM0034	PMED4_00331	conserved hypothetical protein	1.142810855	0.081031866	2.2	0.673031765	14	0.425914356	16.0 Hypothetical	16.1 Conserved hypothetical protein	0.792564	0.057127066	0.8659896	0.02253201	0.7692163	0.13284667
PMM0035	PMED4_00341	DHSS soluble hydrogenase small subunit	1.680248844	0	17.6	0.954801548	5	0.999626228	15.0 Other categories	15.5 Hydrogenase	0.3549726	0.082370184	0.3882134	0.08628413	1.7929748	0.37347206
PMM0036	PMED4_00351	CbiD protein	4.916662016	0.668767818	18.6	0.2798906 17	17	1	2.0 Biosynthesis of cofactors, prosthetic groups, and carriers	2.3 Cobalamin, heme, phycobilin and porphyrin	1.3843483	0.06002715	0.89756	0.12505713	0.89198816	0.24780795
PMM0037	PMED4_00361	guaA Glutamine amidotransferase class-I	1.655972524	0	18.4	0.944146701	6	0.86878047	9.0 Purines, pyrimidines, nucleosides, and nucleotides	9.2 Purine ribonucleotide biosynthesis	0.7345683	0.03384852	0.54932266	0.007214358	1.0550644	0.046346165
PMM0038	PMED4_00371	hypothetical	1.468550903	0.004566636	17.8	0.841759656	6	0.722019453	16.0 Hypothetical	16.2 No similarity	0.90647084	0.03872913	0.6153078	0.023641132	0.8775128	0.05509078
PMM0039	PMED4_00381	conserved hypothetical protein	1.663926057	0	18.6	0.97101744	6	0.689118078	16.0 Hypothetical	16.1 Conserved hypothetical protein	0.3405854	0.019079117	0.34245968	0.019466106	0.74008256	0.19958246
PMM0040	PMED4_00401	Putative penicillin-binding protein	1.466317329	0.004608295	18.2	0.83763254	6	0.838006346	3.0 Cell envelope	3.2 Murein sacculus and peptidoglycan	0.9450296	0.16057572	0.8458101	0.061599165	0.8926737	0.2594661
PMM0041	PMED4_00411	possible reductase	1.625285502	0.000290698	5.2	0.93646514	15	0.842035472	15.0 Other categories	15.7 Other	1.6222777	0.21900965	1.2348988	0.14414412	0.7102872	0.115626566
PMM0042	PMED4_00421	flavoprotein	8.270865769	0.315488086	6.6	0.470853995	18	1	15.0 Other categories	15.7 Other	1.4035434	0.18178836	0.90420467	0.1271419	0.9346152	0.11139351
PMM0043	PMED4_00431	flavoprotein	1.521541682	0.000945626	8.8	0.864090879	2	0.986491182	15.0 Other categories	15.7 Other	1.5163522	0.110555924	1.2928871	0.20083568	1.0979849	0.11144101
PMM0044	PMED4_00441	alaS Alanyl-tRNA synthetase	1.469381056	0.004600939	23.8	0.855104623	10	0.985526907	13.0 Translation	13.1 Aminoacyl tRNA synthetases and tRNA modification	0.947165	0.08724305	0.7520242	0.17790903	0.73643607	0.1953465
PMM0045	PMED4_00451	Orn/DAP/Arg decarboxylases family 2	1.209665721	0.055228031	19.6	0.686164984	8	0.536665341	15.0 Other categories	15.7 Other	0.8790327	0.16487737	0.58193195	0.07572903	0.7269035	0.061833467
PMM0046	PMED4_00461	Nucleoside phosphorylase	1.648961151	0	5.2	0.970899154	16	0.996502995	9.0 Purines, pyrimidines, nucleosides, and nucleotides	9.1 Interconversions and salvage of nucleosides and nucleotides	1.4322665	0.12798996	1.3301909	0.038241312	0.9481033	0.12668619
PMM0047	PMED4_00471	thiO PUTATIVE THIAMINE BIOSYNTHESIS OXIDOREDUCTASE	1.301545812	0.025252525	23.8	0.757948158	10	0.902463702	2.0 Biosynthesis of cofactors, prosthetic groups, and carriers	2.11 Thiamine	0.9115938	0.044635694	0.6603839	0.12434992	0.80907845	0.043093923
PMM0048	PMED4_00481	gatB Glutamyl-tRNA (Gln) amidotransferase subunit B	1.341605269	0.018014956	18	0.766911598	6	0.705255905	13.0 Translation	13.1 Aminoacyl tRNA synthetases and tRNA modification	0.9696149	0.039239094	0.9022607	0.004633606	0.9410464	0.036827195
PMM0049	PMED4_00491	coaE putative dephosph o-CoA kinase	8.859642231	0.260378117	11.8	0.515621255	18	1	10.0 Regulatory functions	10.0 Regulatory functions	1.612187	0.21723436	0.94667697	0.086746074	1.0144494	0.16240361
PMM0050	PMED4_00501	argJ ArgJ family	1.601886472	0.000451467	17.2	0.923649713	5	0.95342144	1.0 Amino acid biosynthesis	1.4 Glutamate family / Nitrogen assimilation	0.9206477	0.12226234	0.8204176	0.03154457	1.1992403	0.19790061
PMM0051	PMED4_00511	conserved hypothetical protein	1.663693315	0	17.6	0.954257108	5	0.999005611	16.0 Hypothetical	16.1 Conserved hypothetical protein	0.5813684	0.08603198	0.46298742	0.07706904	1.4377812	0.48834687
PMM0052	PMED4_00521	Aldo/keto reductase family	8.789877402	0.266306954	19	0.499609241	18	1	15.0 Other categories	15.7 Other	1.0372224	0.09872912	0.7425652	0.2577618	0.83924645	0.09559001
PMM0053	PMED4_00541	Putative RNA methylase family UPF0020	1.382961782	0.01202346	19.2	0.785397217	6	0.769403488	15.0 Other categories	15.7 Other	1.0278219	0.043231465	0.7358064	0.1708444	0.88491136	0.029981917
PMM0054	PMED4_00551	conserved hypothetical protein	1.592432841	0.000417537	17.4	0.914702382	5	0.988756158	16.0 Hypothetical	16.1 Conserved hypothetical protein	0.77814853	0.084418125	0.57607275	0.071438335	1.0636579	0.052226994
PMM0055	PMED4_00561	conserved hypothetical protein	1.566210661	0.000314465	17.6	0.898816271	5	0.941581928	16.0 Hypothetical	16.1 Conserved hypothetical protein	0.4922732	0.041183233	0.3864374	0.11401463	0.7400151	0.24748646
PMM0056	PMED4_00581	conserved hypothetical protein	1.397954246	0.010422961	9.2	0.79546506	1	0.91469533	16.0 Hypothetical	16.1 Conserved hypothetical protein	2.072147	0.013288293	1.5755545	0.060260072	1.9304247	0.37127563
PMM0057	PMED4_00591	putative chromosome segregation protein, SMCATPase superfamily	1.511924839	0.001385681	18	0.867250497	6	0.930164412	13.0 Translation	13.4 Nucleoproteins	0.8993475	0.053966988	0.60225177	0.026606409	0.85613436	0.034305368
PMM0058	PMED4_00601	conserved hypothetical protein	1.616844343	0.000524934	16.6	0.93628503	5	0.783391585	16.0 Hypothetical	16.1 Conserved hypothetical protein	1.0036563	0.058431208	0.8492985	0.17254193	1.1311736	0.10466876
PMM0059	PMED4_00611	conserved hypothetical protein	4.908126225	0.66951567	2.6	0.288924251	18	1	16.0 Hypothetical	16.1 Conserved hypothetical protein	1.3275306	0.35022977	0.7825015	0.058847237	0.7818415	0.36507723
PMM0060	PMED4_00621	accC, fabG acetyl-CoA carboxylase, biotin carboxylasesubunit	1.693407901	0	5.8	0.970932327	16	0.999509191	7.0 Fatty acid, phospholipid and sterol metabolism	7.0 Fatty acid, phospholipid and sterol metabolism	1.6162277	0.016088977	1.3611077	0.17502004	1.0480508	0.14833884
PMM0061	PMED4_00631	YGGT family, conserved hypothetical integralmembrane protein	1.250528588	0.039836925	10	0.715374975	2	0.508754434	15.0 Other categories	15.7 Other	1.5202705	0.39656362	1.0941652	0.3187141	1.0925684	0.006099861
PMM0063	PMED4_00651	conserved hypothetical protein	1.665881259	0	18.8	0.947246729	6	0.963680093	16.0 Hypothetical	16.1 Conserved hypothetical protein	0.70330083	0.07991396	0.50415957	0.024020748	0.84281224	0.210352558
PMM0064	PMED4_00661	hli2 possible high light inducible protein	1.523180815	0.000880503	3.6	0.896390883	12	0.693429233	15.0 Other categories	15.1 Adaptations and atypical conditions	1.2243139	0.064445157	0.8269267	0.016334783	0.7558405	0.020115849
PMM0065	PMED4_00671	ABC transporter, ATP binding protein	1.466815733	0.004616805	1.4	0.863253137	11	0.985007492	14.0 Transport and binding proteins	14.0 Transport and binding proteins	1.1864449	0.22825538	0.76706356	0.004753464	0.76355165	0.07036264
PMM0066	PMED4_00681	conserved hypothetical protein	7.781191191	0.363313388	3.8	0.454721718	18	1	16.0 Hypothetical	16.1 Conserved hypothetical protein	1.2972753	0.2417606	0.82758564	0.07216761	0.7110133	0.121121824
PMM0067	PMED4_00691	HIT (Histidine triad) family protein	1.430918211	0.006990132	23.2	0.827482623	9	0.869731024	10.0 Regulatory functions	10.0 Regulatory functions	1.2865463	0.10230858	0.74297667	0.052366674	0.7035169	0.076632515
PMM0068	PMED4_00701	def putative formylmethionine deformylase	1.437986795	0.006538139	23.8	0.837394559	10	0.958597474	10.0 Regulatory functions	10.0 Regulatory functions	0.902401	0.1226729	0.5470501	0.09776642	0.48494065	0.069403976
PMM0069	PMED4_00711	Esterase/lipase/thioesterase family active site	9.78117275	0.184153713	1.6	0.576095758	18	1	7.0 Fatty acid, phospholipid and sterol metabolism	7.0 Fatty acid, phospholipid and sterol metabolism	1.0401248	0.08557892	0.8791251	0.060330853	0.8783965	0.005547698
PMM0070	PMED4_00721	putative cysteine desulfurase or selenocysteinelyase	9.588501402	0.198429542	2.2	0.564903997	17	1	1.0 Amino acid biosynthesis	1.4 Glutamate family / Nitrogen assimilation	0.9696684	0.093320616	0.7511146	0.12986936	0.7176951	0.055895064
PMM0071	PMED4_00731	ABC transporter, membrane component	1.220428555	0.05186863	2	0.719139188	12	0.571050962	14.0 Transport and binding proteins	14.0 Transport and binding proteins	1.1689254	0.09680775	0.7204035	0.16393468	0.5760259	0.17670156
PMM0072	PMED4_00741	ABC transporter, ATP binding component	1.252977572	0.038807715	1.8	0.738134964	11	0.694283932	14.0 Transport and binding proteins	14.0 Transport and binding proteins	1.275544	0.111973144	0.7436859	0.13889697	0.7753105	0.10289557
PMM0073	PMED4_00751	ABC transporter, membrane component	1.273112504	0.032486388	4	0.742428001	12	0.469165236	14.0 Transport and binding proteins	14.0 Transport and binding proteins	1.1046969	0.043639902	0.82710654	0.17065859	0.7924966	0.09888373
PMM0074	PMED4_00761	conserved hypothetical	7.287563811	0.41251809	19.6	0.41331964	18	1	16.0 Hypothetical	16.1 Conserved hypothetical protein	1.0392781	0.041731276	0.6729787	0.049356334	0.9189953	0.00959389
PMM0075	PMED4_00771	conserved hypothetical protein	1.411994146	0.008658346	16.4	0.882016792	4	0.563699042	16.0 Hypothetical	16.1 Conserved hypothetical protein	1.4211314	0.113625355	1.1678245	0.002356607	1.2629969	0.09121976
PMM0076	PMED4_00791	pgm Phosphoglucomutase	1.607900397	0.000481928	9.4	0.915830722	2	0.990102553	15.0 Other categories	15.7 Other	2.699238	0.07518297	2.3953009	0.050679874	1.6580722	0.067012705
PMM0077	PMED4_00801	yrvN putative ATPase, AAA family	8.741576257	0.270306895	10.8	0.503971165	18	1	10.0 Regulatory functions	10.0 Regulatory functions	1.8188549	0.20547627	1.0234874	0.06462047	1.0174503	0.007597041
PMM0078	PMED4_00811	possible 4'-phosphopantetheinyl transferasefamily protein	1.493839376	0.003131524	3.8	0.87179845	15	0.401284202	15.0 Other categories	15.7 Other	1.413849	0.13229443	0.85692036	0.12160983	0.70698416	0.045314968
PMM0079	PMED4_00821	bcp putative bacterioferritin comigratory (BCP)protein	6.603072551	0.483640153	12.8	0.387522064	17	1	14.0 Transport and binding proteins	14.0 Transport and binding proteins	1.0553485	0.090024845	0.92487985	0.08512669	1.0802752	0.061877467
PMM0080	PMED4_00831	conserved hypothetical protein	1.258354404	0.03696934	4	0.734148484	12	0.542699816	16.0 Hypothetical	16.1 Conserved hypothetical protein	1.4992435	0.016906967	0.77125245	0.15879557	0.7074357	0.003988722
PMM0081	PMED4_00841	cysH Phosphoadenosine phosphosulfate reductase	1.523488947	0.000837321	4.6	0.88365089	15	0.948150494	1.0 Amino acid biosynthesis	1.5 Serine family / Sulfur assimilation	1.3990874	0.11436374	0.83286256	0.3366213	0.49811926	0.036811393
PMM0082	PMED4_00851	putative NADH dehydrogenase, transportassociated	1.057537275	0.131087361	14	0.62314914	18	1	10.0 Regulatory functions	10.0 Regulatory functions	1.1178552	0.20810784	0.8129118	0.112624414	0.9817411	0.070295334
PMM0083	PMED4_00861	putative sodium/sulfate transporter, DASS family	3.964297726	0.760983705	14.2	0.233557227	17	1	10.0 Regulatory functions	10.0 Regulatory functions	1.2260879	0.2850199	1.0164214	0.08557471	1.2188736	0.30367768
PMM0084	PMED4_00871	possible sodium transporter, Trk family	1.365435188	0.014935989	23.4	0.790765271	9	0.756439245	14.0 Transport and binding proteins	14.0 Transport and binding proteins	1.04778	0.08345361	0.7101618	0.06130476	0.8243095	0.039804522
PMM0085	PMED4_00881	putative potassium channel, VIC family	1.550713642	0.00070922	2.8	0.91210292	12	0.840677625	10.0 Regulatory functions	10.0 Regulatory functions	1.0522802	0.014247336	0.82085574	0.06676222	0.6799656	0.10987543
PMM0086	PMED4_00891	Conserved hypothetical protein	1.599432823	0.000443459	20.4	0.907263969	7	0.522993986	16.0 Hypothetical	16.1 Conserved hypothetical protein	0.5500672	0.006034038	0.37825605	0.057531897	0.51147044	0.14324549
PMM0087	PMED4_00901	conserved hypothetical protein	1.670361766	0	20.8	0.948166701	7	0.74466174	16.0 Hypothetical	16.1 Conserved hypothetical protein	0.3149692	0.03922018	0.22569332	0.04990131	0.4384197	0.18751657
PMM0088	PMED4_00911	conserved hypothetical protein	1.658959445	0	20.6	0.941341231	7	0.909267261	16.0 Hypothetical	16.1 Conserved hypothetical protein	0.591259	0.044705473	0.43536025	0.023092587	0.6404928	0.08811097
PMM0089	PMED4_00931	ABC transporter, ATP binding component	1.527986791	0.000857843	22.4	0.876057166	9	0.976865475	14.0 Transport and binding proteins	14.0 Transport and binding proteins	0.93469477	0.006848293	0.62982494	0.049086783	0.7059895	0.028427128
PMM0090	PMED4_00951	possible serine protease	1.460212526	0.004732143	18.4	0.832958453	6	0.862669567	13.0 Translation	13.3 Degradation of proteins, peptides, and glycopeptides	0.7516791	0.23566386	0.47806263	0.288411	0.8474829	0.05797541
PMM0091	PMED4_00961	conserved hypothetical protein	1.367113166	0.0145	16	0.797396603	4	0.851205833	16.0 Hypothetical	16.1 Conserved hypothetical protein	0.7837327	0.27928454	0.77572244	0.37269527	0.7360982	0.2551592
PMM0092	PMED4_00971	conserved hypothetical protein	1.523097279	0.001056338	16.4	0.883323023	4	0.80780674	16.0 Hypothetical	16.1 Conserved hypothetical protein	1.2552583	0.046475135	0.8628687	0.059378903	0.9045016	0.11922276
PMM0093	PMED4_00981	hli1 possible high light inducible protein	1.534066196	0.000899743	8	0.869832781	1	0.958067438	15.0 Other categories	15.1 Adaptations and atypical conditions	1.2798008	0.23096974	1.242 4914	0.02840676	0.8707995	0.114853408
PMM0094	PMED4_00991	conserved hypothetical protein	8.288665424	0.314010778	18.8	0.471407037	17	1	16.0 Hypothetical	16.1 Conserved hypothetical protein	1.1506505	0.116756275	0.8070698	0.17734781	1.0518019	0.19972868
PMM0095	PMED4_01001	similar to serum resistance locus BrkB	3.669775108	0.793333333	2.4	0.216047581	17	1	15.0 Other categories	15.7 Other	1.2881286	0.27398816	0.70925105	0.14871152	0.7571028	0.1481925
PMM0096	PMED4_01011	inositol monophosphate family protein	1.401327313	0.009665145	3.4	0.820696875	12	0.46248625	9.0 Purines, pyrimidines, nucleosides, and nucleotides	9.3 Pyrimidine ribonucleotide biosynthesis	1.2677623	0.050126143	0.90076464	0.08942426	0.837412	0.04648007
PMM0097	PMED4_01021	possible RND family outer membrane effluxprotein	1.568782654	0.000322061	3	0.920955979	12	0.946739878	3.0 Cell envelope	3.1 Membranes, lipoproteins and porins	1.3189116	0.29856354	0.86742324	0.1295197	0.5564758	0.10509581
PMM0098	PMED4_01031	possible Fe-S oxidoreductase	1.360256026	0.015266106	1.8	0.801397527	12	0.572069694	15.0 Other categories	15.7 Other	0.98768157	0.03601639	0.6984192	0.05584322	0.5758181	0.07827693
PMM0099	PMED4_01041	conserved hypothetical protein	4.185649885	0.741788767	17.6	0.240205987	17	1	16.0 Hypothetical	16.1 Conserved hypothetical protein	1.4585186	0.22867443	0.7768081	0.040312942	0.80711263	0.022059591
PMM0100	PMED4_01051	nadB L-aspartate oxidase	6.907806362	0.450651638	14.8	0.406333513	17	1	15.0 Other categories	15.7 Other	1.1886861	0.14107198	0.8999932	0.056382563	1.0231314	0.030020952

Table S2: Characteristics of cell division and (for comparison) cell elongation genes.

Process	PMM number	Gene name(s)	function/ gene product	Peak hour[a]	FDR for periodicity	Cluster	Cluster membership score
Cell division	PMM1309	*ftsZ*	Septal ring	16.6	0.000	5	0.74
	PMM0518	*ftsI, pbp3*	Cell wall biogenesis: cell septation	19.2	0.001	7	0.56
	PMM1458	*ftsW*	cell wall biogenesis: cell septation	19	0.001	7	0.46
	PMM0616	*amiC*	Cell wall hydrolase	3.4	0.034	12	0.69
	PMM0171	*mraW*		N/A	0.139	18 (undetected)	1.00
	PMM0322	*minC*	FtsZ septal ring placement	17	0.031	4	0.53
	PMM0321	*minD*	FtsZ septal ring placement	16.4	0.000	4	0.88
	PMM0320	*minE*	FtsZ septal ring placement	16.8	0.000	5	0.87
Cell elongation	PMM0040	*pbp2*	cell wall biogenesis: cell elongation	18.2	0.005	6	0.84
	PMM1580	*rodA*	cell wall biogenesis: cell elongation	22	0.057	9	0.56

[a] h = 0, is 4 hours after the onset of dark in a 14:10 light-dark cycle.

Table S3: Characteristics of DNA synthesis genes.

Function	PMM number	Gene name(s)	Peak (hour)[a]	FDR for periodicity	Cluster	Cluster membership score
DNA repl. Initiation	PMM0565	*dnaA*	16	0.000	4	0.96
DNA pol III	PMM0001	*dnaN*	18	0.000	5	0.67
	PMM0129	*holB*	N/A	0.196	18 (undetected)	1.00
	PMM0621	*dnaQ*	17	0.000	5	0.81
	PMM0945	*dnaE*	17	0.000	5	1.00
	PMM1658	*dnaX*	19	0.005	6	0.89
Primase	PMM0939	*dnaG*	17	0.000	5	0.89
Helicase	PMM1674	*dnaB*	18	0.003	5	0.73
SS DNA binding protein	PMM1623	*ssb*	17	0.000	5	1.00
DNA ligase (NAD-binding)	PMM0659	*ligA*	19	0.001	6	0.83
DNA ligase (ATP-binding)	PMM0729	*ligB*	N/A	0.812	18 (undetected)	1.00
	PMM1679	*ligB*	N/A	0.137	17 (aperiodic)	1.00
Gyrase (subunit A)	PMM1063	*gyrA*	21	0.001	8	0.50
Gyrase (subunit B)	PMM1634	*gyrB*	16	0.000	4	0.88
Topoisomerase IV (Subunit A)	PMM0005		18	0.000	6	0.63
Topoisomerase I	PMM0436	*topA*	20	0.000	7	0.98
DNA pol I	PMM1140	*polA*	3	0.001	12	0.98

[a] h = 0, is 4 hours after the onset of dark in a 14:10 light-dark cycle.

Table S4: Characteristics of electron transport and ATPase genes.

Function	PMM number	Gene name(s)	function/ gene product	Peak (hour)[a]	FDR for periodicity	Cluster	Cluster membership score
I. Photosynthesis							
PSI	PMM0329	*psaE*		7.4	0.000	1	0.99
	PMM1520	*psaI*		7.8	0.000	1	1.00
	PMM0469	*psaF*		8	0.000	1	1.00
	PMM0540	*psaM*		8	0.000	1	1.00
	PMM1519	*psaL*		8.2	0.000	1	1.00
	PMM1578	*psaD*		8.6	0.000	1	1.00
	PMM0468	*psaJ*		8.6	0.000	1	0.99
	PMM0906	*psaK*		9.4	0.000	1	0.93
	PMM1607	*psaC*		10.2	0.005	1	0.66
	PMM1523	*psaB*	core protein	13.8	0.004	3	0.92
	PMM1524	*psaA*	core protein	14.8	0.004	4	0.87
PSII	PMM1156	*ycf4*		3.6	0.016	12	0.64
	PMM0507	*psb27*		5	0.000	16	0.91
	PMM1098	*psbP*		5.2	0.000	16	0.72
	PMM1152	*ycf37*		7.4	0.000	1	0.91
	PMM0228	*psbO*	manganese stabilizing protein	7.6	0.000	1	1.00
	PMM0926	*psb28*		7.8	0.000	1	0.84
	PMM0251	*psbH*		7.8	0.000	1	1.00
	PMM0272	*psbK*		8.2	0.000	1	1.00
	PMM0253	*psbI*		8.8	0.000	1	0.96
	PMM0317	*psbM*		9.4	0.000	2	0.58
	PMM0299	*psbL*		9.8	0.001	2	0.90
	PMM0314	*psbT*		9.8	0.000	2	0.62
	PMM0315	*psbB*	CP47	10	0.000	1	0.65
	PMM0297	*psbE*		11.2	0.000	3	0.99
	PMM0300	*psbJ*		11.6	0.000	3	1.00

	PMM0298	*psbF*		11.8	0.000	3	1.00
	PMM0223	*psbA*	core protein D1	12.2	0.000	3	1.00
	PMM1158	*psbC*	CP43	12.4	0.000	3	1.00
	PMM0252	*psbN*		12.6	0.000	3	1.00
	PMM1157	*psbD*	core protein D2	12.8	0.000	3	1.00
Other PETC genes	PMM1449	*petF*	ferrodoxin	5.6	0.000	16	1.00
	PMM1171	*isiB*	flavodoxin	8.4	0.000	1	0.98
	PMM1058	*petG*	cytochrome b6f complex	9	0.012	2	0.50
	PMM0581	*petE*	plastocyanin	9.2	0.000	1	0.64
	PMM0627	*pcb*	light harvesting complex	11.4	0.003	3	0.98
	PMM0462	*petC*	cytochrome b6f complex	12.2	0.000	3	1.00
	PMM0461	*petA*	cytochrome b6f complex, cytochrome f	12.8	0.000	3	1.00
	PMM0740	*petN*	cytochrome b6f complex	13.8	0.001	3	0.96
	PMM0325	*petB*	cytochrome b6f complex, cytochrome b6	15.6	0.000	4	1.00
	PMM0326	*petD*	cytochrome b6f complex	15.8	0.000	4	0.99
	PMM1075	*petH*	ferrodoxin-NADP oxidoreductase (FNR)	17	0.095	4	0.42
	PMM1352	*petF*	ferrodoxin	N/A	0.287	17 (aperiodic)	1.00
II. Respiration							
NADH dehydrogenase II	PMM0082	*ndbB*	NDH-2	N/A	0.131	18 (undetected)	1.00
NADH dehydrogenase I	PMM0435	*ndhB*	NDH-1 subunit	1.8	0.002	11	0.86
	PMM1559	*ndhN*	NDH-1 subunit	2.6	0.001	12	0.87
	PMM0594	*ndhD*	NDH-1 subunit	5.6	0.000	16	0.94
	PMM0294	*ndhC*	NDH-1 subunit	18	0.000	5	0.65
	PMM0172	*ndhH*	NDH-1 subunit	18	0.000	5	0.52
	PMM0150	*ndhD*	NDH-1 subunit	18.4	0.000	5	0.92
	PMM0145	*ndhM*	NDH-1 subunit	18.4	0.000	6	0.93

	PMM0293	*ndhK*	NDH-1 subunit	19.2	0.000	6	0.86
	PMM0160	*ndhA*	NDH-1 subunit	19.4	0.000	7	0.69
	PMM0292	*ndhJ*	NDH-1 subunit	19.4	0.001	7	0.55
	PMM0159	*ndhI*	NDH-1 subunit	19.6	0.042	8	0.59
	PMM0570	*ndhL*	NDH-1 subunit	19.8	0.001	7	0.91
	PMM0121	*ndhO*	NDH-1 subunit	20.2	0.070	8	0.41
	PMM0158	*ndhG*	NDH-1 subunit	20.4	0.010	7	0.54
	PMM0157	*ndhE*	NDH-1 subunit	21.4	0.011	9	0.71
	PMM0149	*ndhF*	NDH-1 subunit	N/A	0.986	17 (aperiodic)	1.00
Cytochrome oxidase	PMM0448	*ctaB*	cytochrome oxidase subunit	17.2	0.000	5	1.00
	PMM0447	*ctaA*	cytochrome oxidase subunit	17.4	0.000	5	1.00
	PMM0444	*ctaE*	cytochrome oxidase subunit (III)	18	0.000	5	0.99
	PMM0446	*ctaC (coxB)*	cytochrome oxidase subunit (II)	18	0.000	5	1.00
	PMM0445	*ctaD (coxA)*	cytochrome oxidase subunit (I)	18.2	0.000	5	0.99
III. Proton-translocating ATPase							
	PMM1453	*atpF*	B/B' subunit	5.2	0.000	16	1.00
	PMM1452	*atpH*	delta subunit	5.2	0.000	16	1.00
	PMM1454	*atpG*	B/B' subunit	5.4	0.000	16	1.00
	PMM1455	*atpK*	C subunit	5.4	0.001	16	1.00
	PMM1456	*atpI*	A subunit	5.6	0.000	16	1.00
	PMM1451	*atpA*	alpha subunit	5.8	0.000	16	1.00
	PMM1438	*atpB*	beta subunit	6	0.000	16	1.00
	PMM1450	*atpC*	gamma subunit	6	0.000	16	1.00
	PMM1439	*atpE*	epsilon subunit	6	0.000	16	1.00

[a] h = 0, is 4 hours after the onset of dark in a 14:10 light-dark cycle.

Table S5: Characteristics of the high-light inducible proteins (HLIPs).

PMM Number	Gene name(s)	Peak (hour)[a]	FDR for periodicity	Cluster	Cluster membership score	Copy number
PMM1397/PMM0816[b]	*hli08 / hli18*	1.2	0.074	10	0.65	multi
PMM1398/PMM0817[b]	*hli07 / hli17*	1.4	0.093	10	0.60	multi
PMM1396/PMM0815[b]	*hli09 / hli19*	1.4	0.093	10	0.49	multi
PMM0689	*hli22*	1.4	0.038	11	0.63	multi
PMM1118	*hli04*	1.6	0.001	10	0.52	multi
PMM1399/PMM0818[b]	*hli06 / hli16*	1.8	0.089	10	0.40	multi
PMM1135	*hli14*	1.8	0.020	10	0.41	multi
PMM0064	*hli02*	3.6	0.001	12	0.69	single
PMM0093	*hli01*	8	0.001	1	0.96	single
PMM1317	*hli13*	12.6	0.001	3	1.00	single[c]
PMM1390	*hli10*	14.8	0.000	4	0.99	multi
PMM0471	*hli20*	16.6	0.000	5	0.69	single
PMM1385	*hli11*	21.6	0.000	8	0.82	multi
PMM1384	*hli12*	22	0.001	8	0.83	multi
PMM1482	*hli03*	22.2	0.011	9	0.72	single
PMM1404	*hli05*	N/A	0.155	17 (aperiodic)	1.00	multi
PMM1128	*hli15*	N/A	0.734	17 (aperiodic)	1.00	multi
PMM0690	*hli21*	N/A	0.745	17 (aperiodic)	1.00	multi

[a] h = 0, is 4 hours after the onset of dark in a 14:10 light-dark cycle.
[b] Two identical copies are present in the MED4 genome.
[c] Single copy in many but not all marine cyanobacteria.

Table S6: Characteristics of the carbon metabolism genes.

Pathway	PMM number	Gene name(s)	function/ gene product	Peak (hour)[a]	FDR for periodicity	Cluster	Cluster membership score
Calvin Cycle, glycogen synthesis	PMM0584	*glgB*	1,4-alpha-glucan branching enzyme	4.8	0.000	16	0.94
	PMM0609	*glgA*	ADPglucose--glucosyltransferase	4.8	0.000	15	0.56
	PMM0769	*glgC*	ADP-glucose pyrophosphorylase	5.2	0.000	16	0.98
	PMM0549	*csoS1*	carboxysome	5.2	0.000	16	0.99
	PMM0829	*tpi, cbbJ*	Triosephosphate isomerase	5.4	0.000	16	0.98
	PMM0554	*ccmI, orfA*	carboxysome, putative peptide A	5.4	0.039	15	0.43
	PMM0781	*cbbA, cfxA, fbaA, fda*	Fructose-bisphosphate/sedoheptulose-1,7-bisphosphate aldolase	5.6	0.000	16	0.99
	PMM0767	*glpX, cbbF, fbp*	Fructose-1,6-bisphosphatase/sedoheptulose-1,7-bisphosphatase	5.8	0.001	16	0.98
	PMM0550	*rbcL, cbbL*	carboxysome: Rubisco large chain	5.8	0.000	16	0.99
	PMM0552	*csoS2*	carboxysome	5.8	0.000	16	0.95
	PMM0555	*ccmI, orfB*	carboxysome, putative peptide B	5.8	0.009	15	0.40
	PMM0785	*prk, cbbP*	Phosphoribulokinase	6	0.000	16	1.00
	PMM0551	*rbcS, cbbS*	carboxysome: Rubisco small chain	6	0.000	16	0.98
	PMM0553	*csoS3*	carboxysome: Carbonic anhydrase	6	0.000	16	0.84
	PMM0023	*gap2*	Glyceraldehyde 3-phosphatedehydrogenase	6.8	0.001	16	0.56
	PMM0195	*pgk, cbbK*	Phosphoglycerate kinase	9.4	0.000	2	0.88

Pentose phosphate pathway, glycogen degradation	PMM1322	*glgX*	Glycogen branching enzyme	16.6	0.000	6	0.98
	PMM0519	*tal*	Transaldolase	17.6	0.000	5	1.00
	PMM1601	*glgP*	Glycogen phosphorylase	17.6	0.000	5	1.00
	PMM0770	*gnd*	6-phosphogluconate dehydrogenase	18	0.000	5	0.98
	PMM0771	*devB, pgl*	6-phosphogluconolactonase	18.2	0.000	6	0.84
	PMM1074	*zwf*	Glucose-6-phosphate dehydrogenase	18.8	0.000	6	0.94
Shared use	PMM1489	*rpiA, cbbI*	Ribose 5-phosphate isomerase	4.2	0.000	13	1.00
	PMM0766	*rpe, cbbE*	Ribulose-phosphate 3-epimerase	5	0.000	16	0.99
	PMM1610	*tktA, cbbT*	Transketolase	5.4	0.000	16	1.00
	PMM0076	*pgm*	Phosphoglucomutase	9.4	0.000	2	0.99
	PMM0890	*pgi*	Phosphoglucose isomerase	15.8	0.007	4	0.93

[a] h = 0, is 4 hours after the onset of dark in a 14:10 light-dark cycle.

Table S7: Characterization of the phosphorus metabolism genes.

Categories	PMM number	Gene name(s)	function/ gene product	Peak (hour)[a]	FDR for periodicity	Cluster	Cluster membership score
Regulation	PMM0705	*phoB*	two-component response regulator	19.2	0.005	6	0.80
	PMM0706	*phoR*	two-component sensor histidine kinase	21.8	0.343	18	1.00
Phosphate uptake	PMM0709	*phoE*	porin	21.6	0.000	9	0.91
	PMM0710	*pstS*	ABC transporter, substrate binding protein	3.6	0.030	13	0.67
	PMM0723	*pstC*	ABC transporter, permease component	19.8	0.000	7	0.79
	PMM0724	*pstA*	ABC transporter, permease component	21.6	0.004	9	0.80
	PMM0725	*pstB*	ABC transporter, ATP binding subunit	20.8	0.000	7	0.77
Organic phosphate conversion	PMM0708	*phoA*	alkaline phosphatase	21.8	0.004	8	0.53
	PMM1624	*dedA*	alkaline phosphatase-like protein	14.4	0.026	4	0.72

[a] h = 0, is 4 hours after the onset of dark in a 14:10 light-dark cycle.

Table S8: Characteristics of the nitrogen metabolism genes.

Categories	PMM number	Gene name(s)	function/ gene product	Peak (hour)[a]	FDR for periodicity	Cluster	Cluster membership score
Regulation	PMM0246	*ntcA*	Global nitrogen regulatory protein	N/A	0.647	18 (undetected)	1.00
	PMM0393	*pipX*		17.8	0.000	5	0.97
	PMM1463	*glnB*	Nitrogen regulatory protein P-II	9.4	0.054	1	0.64
Ammonium assimilation	PMM0920	*glnA*	Glutamine synthetase, glutamate--ammonia ligase	16.4	0.000	4	0.81
	PMM1512	*glsF*	Ferredoxin-dependent glutamate synthase,Fd-GOGAT	19	0.000	6	0.95
	PMM1596	*icd*	Isocitrate dehydrogenase	17	0.000	5	0.75
Ammonium uptake	PMM0263	*amt1*	Ammonium transporter	19.2	0.000	6	0.36
Oligopeptide uptake	PMM1049		oligopeptide ABC transporter, substrate binding protein	20	0.000	7	0.92
	PMM0421		putative ABC transporter, oligopeptides	18.8	0.000	6	0.93
Urea uptake	PMM0970	*urtA*	Urea ABC transporter, substrate binding protein	9.6	0.007	1	0.61
	PMM0971	*urtB*	Urea ABC transporter	8.6	0.044	1	0.85
	PMM0972	*urtC*	Urea ABC transporter, membrane protein	N/A	0.261	17 (aperiodic)	1.00
	PMM0973	*urtD*	Urea ABC Transporter, ATP binding subunit	5.6	0.053	15	0.46
	PMM0974	*urtE*	Urea ABC Transporter, ATP binding subunit	3.6	0.054	12	0.67
Urea conversion to ammonium	PMM0965	*ureA*	Urease gamma subunit	19.4	0.000	7	0.62

	PMM0964	*ureB*	Urease beta subunit	20.2	0.000	7	0.92
	PMM0963	*ureC*	Urease alpha subunit	20.6	0.000	7	0.86
	PMM0966	*ureD*	Urease accessory protein	22.6	0.056	9	0.42
	PMM0967	*ureE*	Urease accessory protein	N/A	0.722	18 (undetected)	1.00
	PMM0968	*ureF*	Urease accessory protein	N/A	0.297	18 (undetected)	1.00
	PMM0969	*ureG*	Urease accessory protein	19.8	0.000	7	0.94
Cyanate uptake	PMM0370	*cynA*	Cyanate ABC transporter, substrate binding protein	10.6	0.051	3	0.64
	PMM0371		Cyanate ABC transporter	11	0.269	17 (aperiodic)	1.00
	PMM0372		Cyanate ABC transporter	12.4	0.308	18 (undetected)	1.00
Cyanate conversion to ammonium	PMM0373	*cynS*	Cyanate lyase	11.2	0.019	3	0.91

[a] h = 0, is 4 hours after the onset of dark in a 14:10 light-dark cycle.

Table S9: Characteristics of the regulatory proteins.

Class	PMM number	Gene name(s)	Peak (hour)[a]	FDR for periodicity	Cluster	Cluster membership score
Sigma factors	PMM1629		6	0.000	15	0.54
	PMM1289		14	0.000	4	0.97
	PMM1697		15	0.003	5	0.92
	PMM0577		17	0.000	4	0.52
	PMM0496	*rpoD*	22	0.003	10	0.45
Sensor kinase	PMM0269	*hik01*	20	0.003	7	0.84
	PMM1077	*hik02, sasA*	21	0.062	9	0.71
	PMM1341	*nblS, hik04*	23	0.067	10	0.95
	PMM1579	*hik05*	23	0.043	10	0.91
	PMM0706	*phoR, hik03*	N/A	0.343	18 (undetected)	1.00
Response regulators	PMM0134	*rpaB, rer02*	3	0.000	13	0.96
	PMM1113	*rer05*	10	0.000	1	0.67
	PMM0705	*phoB, rer06*	18	0.005	6	0.80
	PMM0169	*rer03*	23	0.051	8	0.78
	PMM0128	*rpaA, rer01*	N/A	0.141	17 (aperiodic)	1.00
	PMM1619	*rer04*	N/A	0.279	17 (aperiodic)	1.00
Circadian clock	PMM1343	*kaiB*	4	0.000	15	0.86
	PMM1342	*kaiC*	22	0.062	9	0.74
Helix-turn-helix	PMM1637		2	0.040	12	0.86
	PMM1082		3	0.018	10	0.76
	PMM1176		6	0.000	15	0.90
	PMM0154		14	0.009	4	0.78
	PMM0509		16	0.082	6	0.53

	PMM0734		18	0.048	8	0.44
	PMM0988		18	0.000	6	0.52
	PMM1391		18	0.001	6	0.80
fur-like	PMM0637	*fur*	19	0.001	6	0.60
	PMM1030		N/A	0.275	17 (aperiodic)	1.00
crp-like	PMM0718		N/A	0.152	18 (undetected)	1.00
	PMM0806		N/A	0.214	17 (aperiodic)	1.00
N regulation	PMM1463	*glnB, PII*	12	0.054	1	0.64
	PMM0393	*pipX*	18	0.000	5	0.97
	PMM0246	*ntcA*	N/A	0.647	18 (undetected)	1.00
Others	PMM0147	*rbcR, cbbR, lysR*	1	0.005	10	0.49
	PMM1278	*cpmA*	1	0.078	11	0.42
	PMM0363		3	0.021	12	0.82
	PMM0679		3	0.000	15	0.97
	PMM0684		18	0.027	6	0.63
	PMM0939		16	0.000	5	0.89
	PMM1125		5	0.004	15	0.49
	PMM1369		9	0.000	1	0.67
	PMM1642		15	0.023	6	0.59
	PMM0262	*sfsA*	16	0.009	6	0.81
	PMM1262	*lexA*	16	0.000	5	0.96
	PMM0565	*dnaA*	17	0.000	1	0.96
	PMM1393		19	0.059	8	0.39
	PMM0714		23	0.035	8	0.75

[a] h = 0, is 4 hours after the onset of dark in a 14:10 light-dark cycle.

Table S10: Genes with higher or lower expression in continuous light versus the mean expression under a light-dark cycle.

Gene	Function	Continuous: mean diel expression	q-value
PMM0818	hli16 possible high light inducible protein	7.75	7.70E-03
PMM0347	conserved hypothetical	7.19	1.09E-05
PMM0348	possible Spectrin repeat	6.94	4.42E-04
PMM0817	hli17 possible high light inducible protein	6.49	5.37E-03
PMM1396	hli9 possible high light inducible protein	6.25	1.85E-03
PMM0861	possible Virion host shutoff protein	6.10	1.12E-02
PMM1397	hli8 possible high light inducible protein	5.99	1.74E-03
PMM1135	hli14 possible high light inducible protein	5.43	2.87E-03
PMM1629	Type II alternative RNA polymerase sigma factor, sigma-70 family	4.95	7.36E-05
PMM1400	possible Hemagglutinin-neuraminidase	4.83	2.07E-03
PMM1402	Conserved hypothetical protein	4.48	1.88E-05
PMM1390	hli10 possible high light inducible protein	4.37	7.73E-04
PMM1704	dnaK2 Molecular chaperone DnaK2, heat shock protein hsp70-2	4.22	7.72E-05
PMM0699	conserved hypothetical	4.17	2.71E-05
PMM1365	possible MATH domain	4.12	7.84E-06
PMM0901	htpG heat shock protein HtpG	3.68	2.81E-06
PMM0700	conserved hypothetical protein	3.53	4.16E-05
PMM1052	SufE protein probably involved in Fe-S center assembly	3.37	1.99E-03
PMM1028	conserved hypothetical	3.30	1.19E-04
PMM0452	groL GroEL2 protein (Chaperonin cpn60 2)	2.92	4.99E-05
PMM1283	Integral membrane protein, interacts with FtsH	2.64	4.84E-03
PMM0577	Putative type II alternative sigma factor, sigma70 family	2.55	1.49E-04
PMM1118	hli4 possible high light inducible protein	2.53	3.77E-05
PMM1150	putative thioredoxin reductase	2.52	5.68E-05
PMM0016	grpE Heat shock protein GrpE	2.51	1.18E-05
PMM1385	hli11 possible high light inducible protein	2.36	3.95E-04
PMM1405	hypothetical	2.35	4.51E-05

PMM1118	hli4 possible high light inducible protein	2.23	2.92E-06
PMM0407	cysK1 O-acetylserine (thiol)-lyase A	2.23	8.61E-06
PMM0958	conserved hypothetical	2.23	1.30E-04
PMM1289	Type II alternative RNA polymerase sigma factor, sigma-70 family	2.19	2.33E-03
PMM1611	thiC ThiC family	2.15	5.56E-04
PMM1264	ftsH3 cell division protein FtsH3	2.14	3.12E-05
PMM0321	minD putative septum site-determining protein MinD	2.12	8.71E-05
PMM1462	conserved hypothetical protein	2.06	1.58E-02
PMM0690	hli21 possible high light inducible protein	2.04	2.45E-05
PMM1528	HNH endonuclease family protein	2.03	3.06E-04
PMM0043	flavoprotein	2.03	1.08E-03
PMM1437	groES GroES protein (Chaperonin cpn10)	2.01	1.18E-05
PMM0087	conserved hypothetical protein	0.28	7.76E-04
PMM1672	des9 Fatty acid desaturase, type 1	0.32	2.81E-06
PMM1079	possible Villin headpiece domain	0.37	2.81E-06
PMM0305	cpeB Phycobilisome protein	0.38	2.13E-04
PMM1485	rpoB RNA polymerase beta subunit	0.39	8.84E-04
PMM0227	cysD ATP-sulfurylase	0.39	2.92E-06
PMM0751	conserved hypothetical protein	0.42	2.21E-05
PMM0768	hemA glutamyl-tRNA reductase	0.43	1.09E-03
PMM1609	fabF 3-oxoacyl-[acyl-carrier-protein] synthase II	0.45	5.99E-04
PMM0088	conserved hypothetical protein	0.46	2.81E-06
PMM0245	cob(I)alamin adenosyltransferase	0.46	9.40E-05
PMM0399	Putative deoxyribose-phosphate aldolase	0.46	3.56E-04
PMM1501	S1 RNA binding domain:Ribonuclease E and G	0.47	8.89E-05
PMM0496	sigA, rpoD Putative principal RNA polymerase sigma factor	0.47	2.82E-05
PMM0056	conserved hypothetical protein	0.48	6.40E-05
PMM0643	metA putative homoserine O-succinyltransferase	0.48	3.32E-03
PMM1228	hypothetical protein	0.50	7.92E-05

Table S11: Quantitative reverse-transcription-PCR data and primers used.

Gene	Time of Max expression	Time of Min expression	Max/Min expression	Forward primer (5'-> 3')	Reverse primer (5'-> 3')
ftsZ	12:00	08:00	19	ftsZ-675F: AATGACTGAAGCTGGCACTGC	ftsZ-765R: ACTATTCATTGCGGCTTGAGC
dnaA	12:00	00:00	9.4	dnaA-434F: CAGCTTTAGCAGTGGCAGAA	dnaA-538R: AATGACCAACAGCTTGCATC
psbA	12:00	00:00	14	psbA-26F: CTTCGCTGTTAAAAGGCTGGC	psbA-114R: CATTAAGACGCCGAACCAACC
pcb	12:00	00:00	2.6	pcb-90F: TCATGTCGCTCATGCAGGG	pcb-181R: GACCCATTGGGACACTGGG
rbcL	04:00	16:00	53	rbcL-55F: CCTGAATATGTCCCCCTCGA	rbcL-145R: CCGCTGCAACTTCTTCT
kaiC	20:00	12:00	3.7	kaiC-730F: GCCTTAGGAGCGATGAGATT	kaiC-825R: AAAATAACCCCCTCCACACA
rpoD	20:00	12:00	17	rpoD-591F AATCAGAGCTGCCGAAAAAT	rpoD-692R TGATCTGCTATCGCTCGTGT
rnpB	N/A	N/A	N/A	rnpB-1F: TTGAGGAAAGTCCGGGCTC	rnpB-91R: GCGGTATGTTTCTGTGGCACT

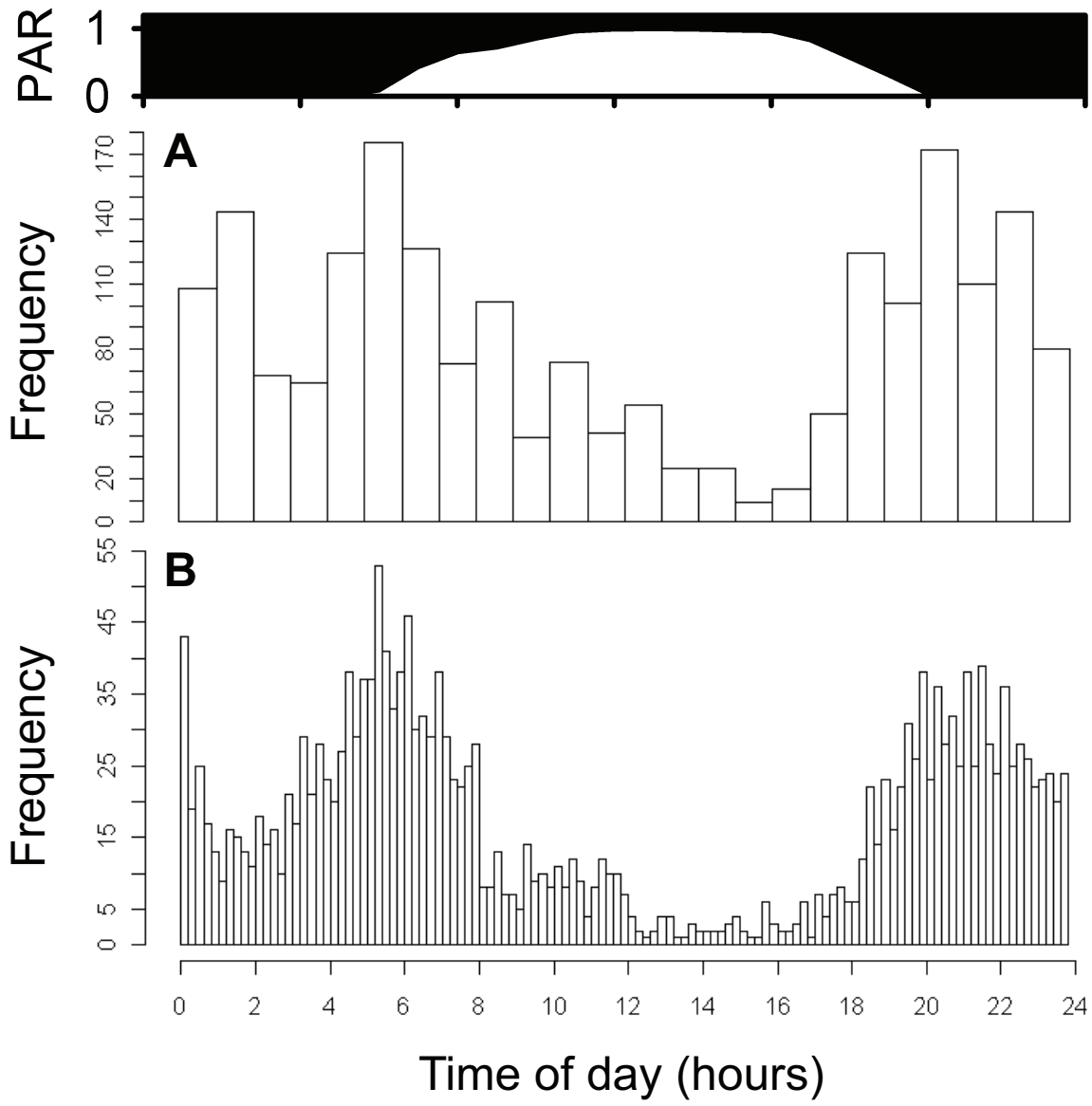

PAR
1
0
A
Frequency
0
20
50
80
110
140
170
B
Frequency
0
5
15
25
35
45
55
0
2
4
6
8
10
12
14
16
18
20
22
24
Time of day (hours)

Figure S1

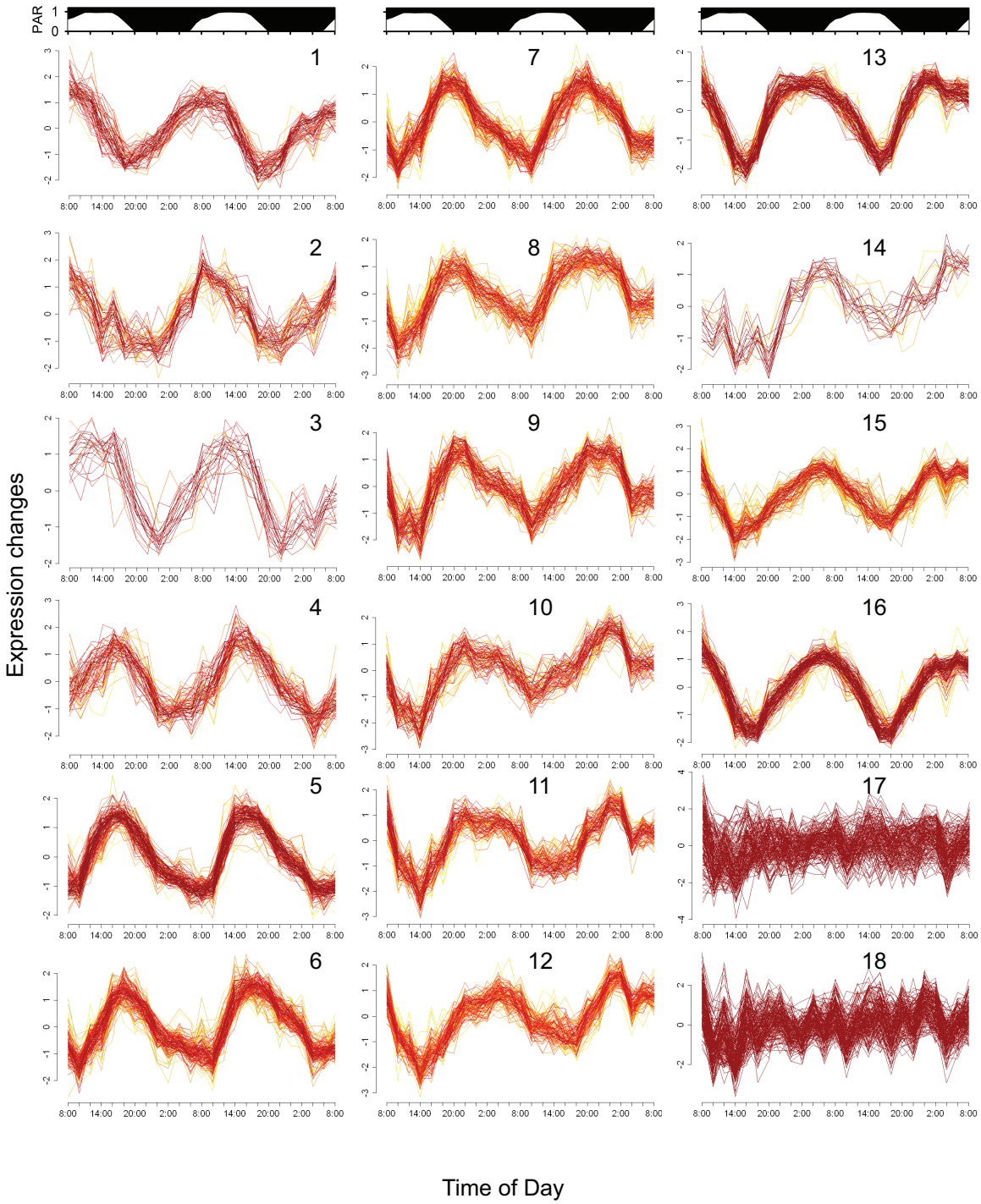

PAR
1
0
1
7
13
2
8
14
3
9
15
4
10
16
5
11
17
6
12
18
Expression changes
Time of Day

Figure S2

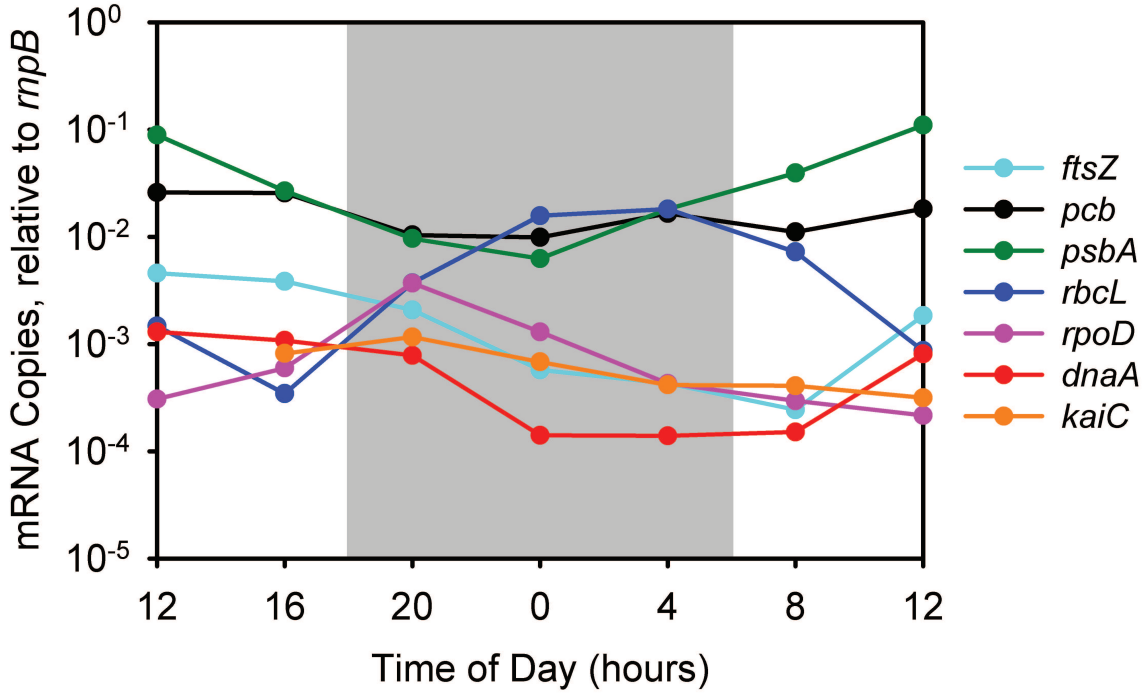

mRNA Copies, relative to rnpB
10^0
10^-1
10^-2
10^-3
10^-4
10^-5
12
16
20
0
4
8
12
Time of Day (hours)
ftsZ
pcb
psbA
rbcL
rpoD
dnaA
kaiC

Figure S3

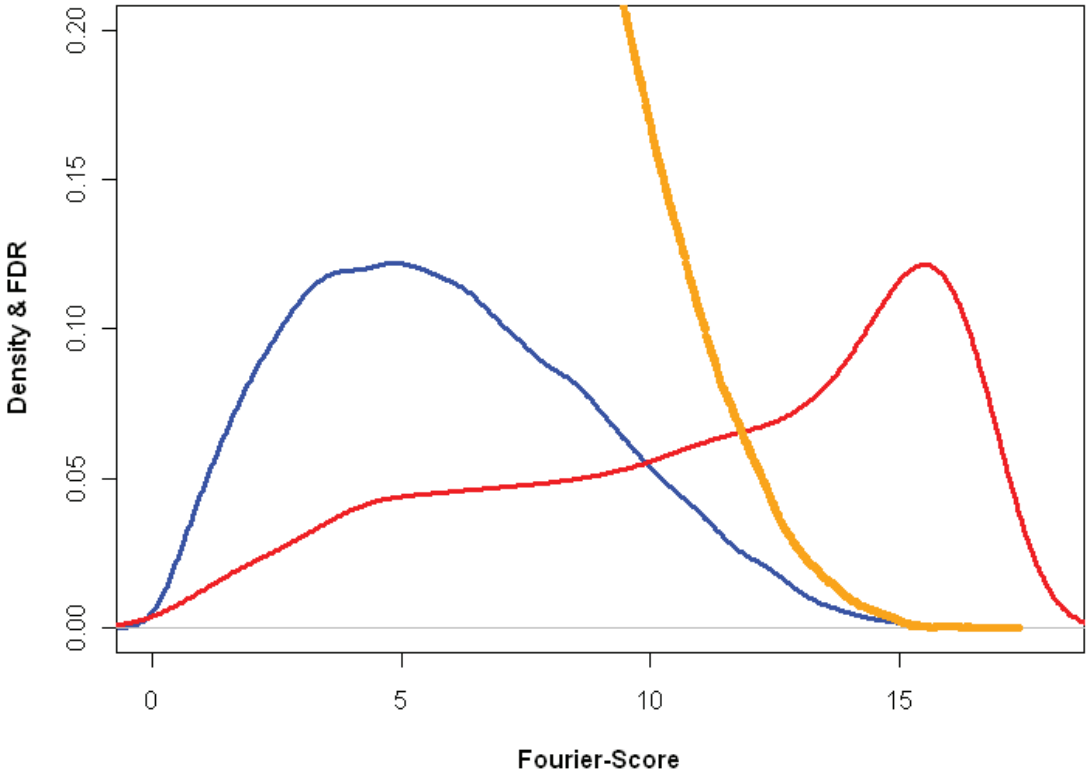


Figure S4

APPENDIX G

Genomic analysis of oceanic cyanobacterial myoviruses compared to T4-like myoviruses from diverse hosts and environments

(Sullivan et al., *Environ Microbiol*, in press)

Genomic analysis of oceanic cyanobacterial myoviruses compared to T4-like myoviruses from diverse hosts and environments

Matthew B. Sullivan*[,1, 2], Katherine H. Huang[1] , Julio C. Ignacio-Espinoza[2], Aaron M. Berlin[3], Libusha Kelly[1], Peter R. Weigele[1, 4], Alicia S. DeFrancesco[1], Suzanne E. Kern[1], Luke R. Thompson[1], Sarah Young[3], Chandri Yandava[3], Ross Fu[1], Bryan Krastins[5], Michael Chase[5], David Sarracino[5], Marcia S. Osburne[1], Matthew R. Henn[3], Sallie W. Chisholm*[,1]

[1] Massachusetts Institute of Technology, Cambridge, MA USA, [2] University of Arizona, Tucson, AZ USA, [3] Broad Institute, Cambridge MA USA, [4] New England Biolabs, Chemical Biology Division, 240 County Road, Ipswich, MA 01938, [5] Harvard Partners, Cambridge MA 02139

* co-corresponding authors: Matthew Sullivan (mbsulli@email.arizona.edu) + Sallie Chisholm (chisholm@mit.edu)

ABSTRACT: T4-like myoviruses are ubiquitous, and their genes are among the most abundant documented in ocean systems. Here we compare 26 T4-like genomes including 10 from non-cyanobacterial myoviruses, and 16 from marine cyanobacterial myoviruses (cyanophages) isolated on diverse *Prochlorococcus* or *Synechococcus* hosts. A core genome of 38 virion construction and DNA replication genes was observed in all 26 genomes, with 32 and 25 additional genes shared among the non-cyanophage and cyanophage subsets, respectively. These hierarchical cores are highly syntenic across the genomes, and sampled to saturation. The 25 cyanophage core genes include 6 previously described genes with putative functions (*psbA, mazG, phoH, hsp20, hli03, cobS*), a newly described phytanoyl-CoA dioxygenase, 2 virion structural genes, and 16 hypothetical genes. Beyond previously described cyanophage-encoded photosynthesis and phosphate stress genes, we observe here core genes that likely play a role in nitrogen metabolism during infection through modulation of 2-oxoglutarate. Patterns among non-core genes that may drive niche diversification revealed that phosphorus-related gene content reflects source waters rather than host strain used for isolation, and that carbon metabolism genes appear associated with putative mobile elements. As well, phages isolated on *Synechococcus* had higher genome-wide %G+C and often contained different gene subsets (e.g., *petE*, *zwf*, *gnd, prnA, cpeT*) than those isolated on *Prochlorococcus.* However, no clear diagnostic genes emerged to distinguish these phage groups, suggesting blurred boundaries due to cross-infection. Finally, genome-wide comparisons of both diverse and closely-related, co-isolated genomes provide a locus-to-locus variability metric that will prove valuable for interpreting metagenomic datasets.

INTRODUCTION

T4-like phages

Double-stranded DNA bacteriophages (Caudovirales) are the primary viral types observed in marine systems. Myoviruses (contractile-tailed phages) predominate among these, as determined by viral metagenomic surveys (Breitbart et al., 2002; Breitbart et al., 2004; Angly et al., 2006; Williamson et al., 2008) and in culture experiments (Suttle and Chan, 1993; Waterbury and Valois, 1993; Wilson et al., 1993; Lu et al., 2001; Marston and Sallee, 2003; Sullivan et al., 2003). Myoviruses also dominated the viral signal in *microbial-fraction* metagenomic datasets from Hawaii (DeLong et al., 2006) and from the surface waters sampled in the Global Ocean Survey (GOS, (Rusch et al., 2007; Yooseph et al., 2007); the latter of which reports that 5 of the 6 most abundant GOS proteins were attributed to T4-like myoviruses (Yooseph et al., 2007). The viral signal in these microbial metagenomes is thought to represent infecting viruses captured inside infected host cells, suggesting that T4-like phages are both numerically abundant and actively infectious (DeLong et al., 2006).

The canonical *E. coli* bacteriophageT4 has a well-characterized infection cycle, genome and transcriptome (Luke et al., 2002; Miller et al., 2003a). A watershed of papers has defined the “core” genes representative of the growing family of known T4-like phages. Relatively early work (Hambly et al., 2001) first noted that the ocean cyanobacterial T4-like virus S-PM2 had a module of capsid gene sequences similar to those of phage T4 - isolated using *Escherichia coli* from sewage - suggesting that at least

portions of these phage genomes might be shared across distantly related phages. Subsequent work (Desplats et al., 2002) expanded these observations, using a larger fraction of an *E. coli* T4-like phage genome (RB49) to show that the general virion structural components *and* the DNA replication apparatus were also conserved across T4-like phages. Whole genome comparison followed that compared the archetype T4 phage to marine T4-like vibriophage KVP40 (Miller et al., 2003b), and T4-like coliphage JS98 (Chibani-Chennoufi et al., 2004); these studies showed that the "T4 core" genes encode structural proteins to produce virus particles, as well as the metabolic machinery required for infection of the host.

As new genomes became available, further whole genome comparisons refined our understanding of the T4 core (e.g., phages T4, RB49, and Aeh1 share 90 genes, Comeau et al., 2007) and shifted the focus to characterizing the flexible genome of T4-like phages (Nolan et al., 2006). These flexible genes encode proteins that interact with the host cell, e.g., tail fibers and internal scaffolding proteins, or likely offer other niche-defining functions such as base modification and differential complements of tRNAs (Comeau et al., 2007). Most of these genes are thought to represent ancient lateral transfer events, as 90% of them exhibited early/middle promoter control similar to that seen for the corresponding T4 core genes (Nolan et al., 2006).

Cyanobacterial T4-like phages

Ocean microbes drive globally-important biogeochemical cycles, including carbon, oxygen, nitrogen, and sulfur cycles (Arrigo, 2005; Howard et al., 2006; Karl, 2007), and the enormous numbers of ocean viruses (typically $>10^7$ ml^{-1}, or approximately ten for every microbial cell) drive the evolution of microbial processes through host mortality (Fuhrman, 1999; Wommack and Colwell, 2000; Weinbauer, 2004; Suttle, 2005), horizontal gene transfer (Paul, 1999; Miller, 2001), and the modulation of host metabolism (Breitbart et al., 2007). Among marine microbes, the picocyanobacteria *Prochlorococcus* and *Synechococcus* are highly abundant (Waterbury et al., 1979; Waterbury et al., 1986; Partensky et al., 1999), and some estimates suggest that they account for as much as one-third of oceanic primary production (Li, 1994; Li, 1995). These two genera are commonly present at 10^5 cells ml^{-1} and usually co-occur: *Prochlorococcus* is numerically dominant in the vast, low nutrient open oceans (Partensky et al., 1999; Johnson et al., 2006; Coleman and Chisholm, 2007), while *Synechococcus* dominates in coastal waters (Waterbury et al., 1979; Waterbury et al., 1986).

In previous studies, four *Prochlorococcus* and *Synechococcus* T4-like cyanophage genomes were found to share up to 45 genes (out of ~150 total) with the non-cyanophages (Mann et al., 2005; Sullivan et al., 2005; Weigele et al., 2007, but also see Millard et al. 2009 and "note in proof"). In addition, these studies revealed the power of phage–host co-evolution in the context of ocean-basin scale ecological settings. For example, cyanophage genomes were found to contain "host" genes involved in central host metabolism and photosynthesis (Mann et al., 2003; Lindell et al., 2004; Millard et al., 2004, 2009; Mann et al., 2005; Sullivan et al., 2005; Weigele et al., 2007), and these genes are expressed during phage infection (Lindell et al., 2005; Clokie et al., 2006; Lindell et al., 2007). Further, the viral version of these host genes dominates the GOS surface ocean *microbial-fraction* metagenomes, e.g., 60% of the identifiable *psbA* genes were viral (Sharon et al., 2007). The distributions of these host photosynthetic genes among phage types appear driven by the physiology of the phage (e.g., host range for *psbD* and lytic cycle length for *psbA,* Sullivan et al., 2006). In fact cyanophages may be among the drivers of photosystem evolution as portions of the "host" genes carried on cyanophages are able to recombine back into the host gene pool (Zeidner et al., 2005; Sullivan et al., 2006).

In contrast to the near-ubiquity of the core photosystem II *psbA* gene present in cyanophage genomes, other "host" genes are sporadically distributed among cyanophage genomes but also may impact phage fitness. On the one hand, T4-like viral contigs assembled from marine metagenomes contain up to seven clustered photosystem I genes thought to form an intact monomeric PSI complex to funnel reducing power from electron transport chains to PSI-related functions during infection (Sharon et al., 2009). Interestingly, such PSI genes have yet to be identified in any genome from a cyanophage isolate (Chen et al. 2002, Mann et al. 2005, Millard et al. 2009, Sullivan et al. 2005, Weigele et al. 2007). On the other hand, the functional role of cyanophage-encoded phycobilin synthesis genes (*pcyA* and *pebS*) remains a mystery (Dammeyer et al., 2008). In this case, despite the fact that *Prochlorococcus* hosts lack intact phycobilisomes and that these cyanophage-encoded genes are highly divergent relative to host copies, they are expressed *in vivo* during infection and are functional *in vitro* (Dammeyer et al., 2008). It is likely that these and other sporadically distributed genes serve specific niche-defining roles for

phages' adaptation to their particular hosts and environments that will reveal themselves as more genome and physiology data become available.

Here we expand the T4-like cyanophage database, nearly doubling the number of T4-like phage genomes by adding 12 new ocean cyanophage genomes to the previous 4 (Table 1). We use this augmented database to explore the ecology and evolution of T4-like cyanophages through an analysis of the genomes of 16 marine cyanophages compared with 10 non-cyanophage T4-like genomes from the Tulane Genome Sequencing Project (http://phage.bioc.tulane.edu/). The cyanophages were isolated from 15 different habitats over a period of 16 years, using 10 different host strains (4 *Prochlorococcus* and 6 *Synechococcus*), while the non-cyanophages were isolated over decades using at least 7 different source waters and 6 different hosts. Thus, these conditions optimize the potential for revealing diversity across the 26 phage isolates (Table 1) examined in this study. With this dataset, we asked the following questions: What gene sets are shared and not shared among various hierarchical groupings of T4-like phages, and how do these genes inform our understanding of T4-like cyanophage and non-cyanophage biology? What mechanisms likely drive differential and sporadic distribution of non-shared genes among the cyanophages?

RESULTS AND DISCUSSION

General features of the sixteen cyanophage genomes

All available annotation information for the 16 cyanophage genomes is provided in a detailed overview figure (Fig. 1). With two exceptions, cyanophage genome sizes ranged from 174-196kb (summarized in Table 1, details provided in Suppl. Table 1), as commonly observed previously for non-cyanophages (Miller et al., 2003b; Miller et al., 2003a; Nolan et al., 2006; Petrov et al., 2006). The exceptional cyanophages were S-SSM7 (232kb) and P-SSM2 (252kb), which contained large lipopolysaccharide gene clusters (Fig. 1, discussed below) that accounted for about 72-85% of the expanded genome size. Cyanophage genome size was correlated with the number of predicted ORFs (R^2 = 0.743), and there was no apparent relationship between the genome size and the genus of the host on which it was isolated (Suppl. Fig. 1).

While significant variation in genome-wide %G+C exists among the non-cyanophages (Table 1), even for those isolated on the same host, we note that this metric is less variable among the cyanophage genomes (Table 1). As well, the average genome-wide %G+C content of phages isolated on *Prochlorococcus* (37.2±1.0%) is significantly different (P≤0.0001) from that of phages isolated on *Synechococcus* (40.1±1.0%). Such cyanophage variability may reflect host-range constrained swapping of genetic material followed by subsequent genome-wide amelioration of the new genes in the phage genome. For example, *Synechococcus* cells have higher %G+C genomes than *Prochlorococcus* (Kettler et al., 2007; Dufresne et al., 2008) and even high %G+C material from *Synechococcus* hosts would ameliorate once in the phage genome towards the overall lower %G+C of phage genomes. In contrast, *Prochlorococcus* phage %G+Cs are often closer to that of their host genomes, so the impact of such genome-wide amelioration pressures are minimal compared to that seen in *Synechococcus*. Such observations in cyanophage-encoded core photosynthesis genes proved diagnostic for tracing intragenic recombination events among cyanophage genomes (Zeidner et al., 2005; Sullivan et al., 2006). That one cyanophage, S-PM2, deviates from the general pattern may hold clues regarding the host range of this particular phage (also see below).

Gene distributions among hierarchical groupings of the genomes

In preparation for analyses of gene content and order in the different genomes, we clustered orthologous genes into T4 *G*ene *C*lusters ("T4-GCs"; see methods), and used these to define core gene sets common to hierarchical groupings of the genomes (Fig. 2a, see discussion below). A total of 7,071 predicted genes in the 26 genomes clustered into 874 T4-GCs, with 1,941 genes remaining as singletons.

Gene presence/absence network analysis: To examine how similar the genomes are to each other with respect to the presence or absence of each T4-GC, we represented the presence/absence table as a network (Fig. 2b), that links T4-GCs to the genomes in which they are found. Genomes with many T4-GCs in common appear in close proximity due to the many connections that they share. The resulting network shows clustering of the cyanophage (blue diamonds, Fig. 2b) separate from non-cyanophage (red diamonds, Fig. 2b) T4-like genomes by this metric. Core genes shared by all 26

genomes connect the two groups of phage and are highlighted as the single, central purple circle (Fig. 2b).

Core and pan-genomes: To explore the features of the core and pan-genomes of the cyanophage and non-cyanophage subsets given the number of genomes sequenced, we identified the shared and unshared gene sets of all possible combinations of choosing k genomes (k = 1 to n) from n sequenced genomes (Fig 3).The core genes shared within the two groups (discussed in detail below) leveled off quickly as new genomes were added to the analysis, suggesting that this small sample size of diverse T4-like phages is adequate for determining the core. As expected, the total number of unique genes identified (the pan-genome) steadily increased with the number of available genomes in both cases. The size of the pan genome reached 1,422 and 1,445 genes for the cyanophages and non-cyanophages, respectively (Fig. 3a and 3b). The rate of increase of both pan genomes as more genomes are added to the analysis is far from saturated, indicating the existence of a much larger and diverse gene pool than has been captured by the 26 sequenced genomes. Interestingly, however, the cyanophage pan genome showed a slower rate of increase (Fig. 3a) than that of the non-cyanophages (Fig. 3b).

The T4 core, shared by all 26 T4-like phage genomes: Thirty-eight genes were common to all 26 genomes (Fig. 2a, Suppl. Table 2), while also maintaining remarkable synteny (Suppl. Fig. 2). The only exceptions to the synteny included a large inversion among the cyanophages relative to the non-cyanophages, and a few notable smaller-scale breaks in synteny likely due to mobile element activity (see the "genomic evolution" section). Of the 38 genes shared by all the genomes, 27 form sequence-based orthologous groups (T4-GCs; see methods), while the remaining 11 display enough sequence divergence that these functional homologs are placed into multiple T4-GCs. While the number of core genes decreased as more T4-like phage genomes were added to these analyses (Miller et al., 2003b; Mann et al., 2005; Sullivan et al., 2005; Comeau et al., 2007; Weigele et al., 2007; Millard et al., 2009), it appears that we have now adequately defined the core (Fig. 3) and that these*T4 core* functions involve appropriating host metabolic machinery, replicating the viral genome during infection, and building the viral particles.

Nearly "T4 core" genes: Beyond the *T4 core* genes are a handful of noteworthy *nearly core* genes, i.e., those present in at least 22 genomes across the 26 T4-like phage genomes. An analysis of the patterns of their distributions makes these genes potentially useful targets for experimental functional identification, or indicators of novel functions in particular groups of isolates. This set of genes includes the gp51 baseplate hub assembly catalyst (missing only in Aeh1, but note that cyanophage gp51 are only ~20% of the length of non-cyanophage gp51, Suppl. Fig. 3), nucleotide metabolism and recombination / repair genes *uvsX, uvsY* (both missing in the same three phages – 44RR, PHG25, PHG31), and the gp59 loader of gp41 helicase (found in 22 of 26 T4-like phages).

The non-cyanophage core, shared by all 10 non-cyanophage genomes: In addition to the 38 genes shared by all the genomes, the non-cyanophage genomes shared an additional 32 *non-cyanophage core* genes (Fig. 2a, Suppl. Table 3), giving this group a shared core of 70 genes down from the most recent estimate of 90 core genes shared among 3 non-cyanophage T4-like genomes (Comeau et al. 2007). All but 6 of the 32 *non-cyanophage core* genes have been functionally annotated in coliphage T4 (Miller et al., 2003a), and the larger proteins such as structural proteins gp7, gp10, and gp12 were so divergent as to be comprised of up to 9 T4-GC clusters (Suppl. Table 3). Many of these additional *non-cyanophage core* genes encode functions involved in "host specialized" viral structure (e.g., tail fibers) and DNA replication machinery. We expect that experiments targeting functional annotation of shared hypothetical proteins in the cyanophages will reveal that many of these host specific functions exist in the cyanophages, but as divergent gene copies. In contrast, other genes, such as *nrdD* and *nrdH* genes, are likely only relevant to the specific habitat of some of these non-cyanophages (e.g., anaerobic sewage).

The cyanophage core, shared by all 16 cyanophage genomes: Twenty-five genes were shared by all 16 cyanophages (Fig. 2a, Suppl. Table 4), in addition to the 38 that form the *T4 core*, for a total of 63 genes shared across the cyanophages which now appears to be a stable shared gene set among the T4-like cyanophages (Fig. 3a). All but one of these 25 *cyanophage core* genes was absent from the non-cyanophages (Suppl. Table 4). This exception is the *phoH* gene that was found in only one of the other genomes – the marine vibriophage KVP40 – and may represent an adaptation valuable both for infection of cyanobacteria, but also more generally of marine hosts (e.g., marine vibrios) rather than a cyanophage-specific function. However, some do appear cyanophage-specific, such as the previously described cobalamin biosynthesis protein (*cobS)*, or photosynthesis proteins for the central photosystem

II reaction center protein (*psbA*) and high-light inducible proteins (*hli*) (Mann et al., 2005; Sullivan et al., 2005; Weigele et al., 2007). Other *cyanophage core* genes include proteins that likely encode basic phage functions, such as a heat shock family protein (*hsp20)* that might be important for scaffolding during maturation of the capsid*,* and 2 experimentally determined virion structural proteins (T4-GCs 15, 190). In addition, the *cyanophage core* includes, phosphate-stress induced protein (*phoH*)*,* pyrophosphatase (*mazG),* and dioxygenase proteins (T4-GCs 101, 155 with similarity to PFAM PF05721) that are discussed in greater detail below. The remaining genes encode hypothetical proteins of unknown function. An understanding of the functions of these proteins, combined with a deeper understanding of the PhoH and MazG proteins (discussed below) should further elucidate the nature of cyanophage–host interactions.

Notable cyanophage core and nearly cyanophage core genes: The *cyanophage core* gene *mazG* has received a lot of recent attention. In *E. coli,* MazG appears to be a regulator of nutrient stress and programmed cell death (Magnusson et al., 2005; Gross et al., 2006; Lee et al., 2008), as its dNTP pyrophosphatase activity acts on the signaling nucleotide guanosine tetraphosphate (ppGpp) to regulate up to 1/3 of *E. coli* genome (Traxler et al., 2008). In cyanophages, MazG is also thought to act as a global transcriptional regulator through modulation of ppGpp levels, which may extend the period of cell survival under the stress of phage infection (Clokie and Mann, 2006; Weigele et al., 2007). However, MazG enzymes are highly specific for non-canonical NTPs, suggesting that identifying their substrates likely requires solving crystal structures along with activity and binding assays for each new enzyme (Galperin et al. 2006). Thus the cyanophage MazG substrate should be cautiously interpreted.

Regardless of function, the *mazG* gene has a notable distribution among T4-like cyanophages. Recently, it was found by PCR screens to be present in 9 out of 17 cyanophage myovirus isolates (Bryan et al., 2008). In contrast, all 16 of our cyanophage myovirus genomes contained this gene. While this difference could be real, it likely reflects the limitations of PCR screening, which can only reveal the presence (not the absence) of a gene in a particular genome with confidence because primers can only be designed to capture known sequences (Millard et al., 2004; Millard et al., 2009). Consistent with this interpretation, Bryan et al. (2008) observed >99% identity among their sequenced *mazG* PCR products obtained from geographically diverse isolates, while the *mazG* sequences of our genomes showed marked sequence divergence (Suppl. Fig. 4). Nonetheless, in agreement with Bryan et al. (2008), our analyses also suggest that *mazG* arose from outside the cyanobacteria (Suppl. Fig. 4), as opposed to most other "host" genes in cyanophages which originate from their host strains (Sullivan et al., 2005; Williamson et al., 2008), and is most closely related to the genes from *Choloroflexus*.

Finally, in addition to the core *mazG* gene, nine genes are *nearly cyanophage core* genes as they are found in 15 of the 16 cyanophages, missing only in the anomalous S-PM2 phage (see below).

Genome variability of two co-isolated cyanophages: To explore genomic diversity among spatially co-existing phages capable of infecting the same host, we included in this sequencing project two phages isolated from the same water sample on the same host strain (Fig. 4a). These two cyanophages, P-HM1 and P-HM2, are highly syntenic and share 200 of 246 genes, whose protein sequences are on average 83% identical (Fig. 4a). In contrast, pairwise genome comparisons showed that among the non-co-isolated cyanophages, the genomes share as much as 77-80% of their genes with average identity 72-75% (Fig. 4b) or at the least 22-33% of their genes, with only 48-49% average identity (Fig. 4c).

Further comparison of the two co-isolated phage genomes (Fig.4a) showed that, while the protein identity of orthologs shared between P-HM1 and P-HM2 averaged 83%, there was an enormous range (21-100%) in this value. On the one hand, ortholog identities could be quite low (21-32%) and include hypothetical proteins and even proteins that are part of the *cyanophage core* such as CoA-dioxygenase and Hsp20. On the other hand, ortholog identities could be quite high (100%) for other *cyanophage core* proteins such as Hli03, gp55, as well as for non-conserved hypothetical proteins such as T4-GCs 429, 542, and 559, which are found only in a sub-set of *Prochlorococcus* phages. The non-shared proteins, predominately hypotheticals, were notably clustered into distinct regions of the genomes (Fig. 4a) akin to cyanobacterial genomic 'islands' (*sensu* Coleman et al. 2006). In addition to hypotheticals, the non-shared gene set did include some annotation (Suppl. Table 5): a purine biosynthesis gene (*purM*) and plastoquinol terminal oxidase (PTOX, described further below) are unique to P-HM1, while a pair of endonucleases and a Kelch-repeat-containing protein are unique to P-HM2. In addition, peptidase genes were present in syntenic genomic locations in both phages (Fig. 4A) even though their sequences had diverged to the point of forming separate gene clusters (T4-GCs 573, 452). These phages also contain 66 genes found in both phages, but not in any of the other sequenced cyanophages. These 66 genes

encode an S8 peptidase (T4-GC518), glycine dehyrogenase (T4-GC540), two asparaginyl beta-hydroxylases (T4-GCs 536, 546), an acyl carrier protein (ACP, T4-GC457) and its synthetase (ACPS, T4-GC500), a terminal quinol oxidase (T4-GC555), taurine catabolism dioxygenase (T4-GC447), and hypotheticals. That genes encoding these proteins were found only in these two co-isolated MED4-infecting phages might provide clues to requirements for infection of *Prochlorococcus* MED4 in these Hawaii Ocean waters.

The cyanophage-exclusive, but not universal, gene set: We identified 143 genes that occurred in 4 or more of the 16 cyanophage genomes, but were absent from all of the non-cyanophage genomes (summarized in Table 2). Ninety-six of these encode hypothetical proteins, but others encode a diversity of photosynthesis (*psbD, petE, petF,* PTOX*, pebS*), phosphate stress (*pstS*), carbon metabolism (*talC*, CP12), and virion structural (24 genes) proteins, the functions of which are consistent with our notion of a cyanophage lifestyle. Some of these are discussed further below.

The Synechococcus-enriched gene set: We found no genes that were universal and exclusive to the 10 cyanophages isolated on *Synechococcus*. However*,* there were 48 genes that occurred in 3 or more of this phage set, and occurred in no others (Suppl. Table 6). Notably, these genes clustered in four "hot-spot" regions of the genomes: (a) near gp5 with tRNAs, (b) with small genes between gp46 and gp25, (c) between gp16 and gp17 (previously identified by Millard et al. 2009), and (d) near *psbA*, again commonly with numerous tRNAs (Fig. 1). Although 42 of these 48 genes encode hypothetical proteins, two are involved in carbon metabolism (*zwf, gnd* – discussed below), three had PFAM domains that suggested function (PA14 carbohydrate binding domain, DUF1583, and SAICAR synthetase purine biosynthesis), and one is a virion structural protein (T4-GC969; see "Experimentally identified cyanophage structural proteins").

The Prochlococcocus T4 core and enriched gene set: Two genes were universal and exclusive to cyanophages isolated on *Prochlorococcus* (Suppl. Table 7). These *Prochlorococcus* T4 core genes encode a possible photosystem II PsbN (Pfam domain PF02468, T4-GC163, no functional role has yet been determined for PsbN), and a hypothetical (T4-GC285). As well, there were 16 more genes that occurred in 3 or more of this phage set, and occurred in no others (Suppl. Table 7). These clustered in "hot-spot" genome regions homologous to those described above for the *Synechococcus* enriched genes (Fig. 1), and include genes encoding a high-light inducible protein (T4-GC436), a phycocyanobilin biosynthesis protein (*pcyA*, T4-GC413), and 14 hypothetical proteins. Finally, two hypothetical proteins were universal among the 6 *Prochlorococcus* phages, but not exclusive to them (T4-GC082 also found in S-SSM7 and S-SSM5; T4-GC224 also found in S-SSM7).

The odd cyanophage out: Synechococcus cyanophage S-PM2 appears quite distinct from the 15 other cyanophages. First, its %G+C content is similar to that of a *Prochlorococcus* phage (Table 1). Second, S-PM2 lacks nine *nearly cyanophage core* genes that are found in all of the 15 other cyanophages, and two genes found in 14 of the 15 other cyanophages. In contrast, only one other cyanophage (P-SSM2) is missing even a single gene (T4-GC424) that is ubiquitous among the other 15 cyanophages. Among the genes "missing" in S-PM2 are 8 hypothetical genes, an endonuclease, and two carbon metabolic proteins (transaldolase and CP12 = T4-GCs 63, 337). Finally, S-PM2 contains only seven of the 45 "*Synechococcus*-enriched*"* phage genes, whereas, other than *Synechococcus* phage S-SSM7 (containing only two), the rest of the cyanophage genomes contained 18–27 (average = 23) of the 45 *Synechococcus* phage-enriched genes. Given the data set at hand, we cannot identify any variables that might explain why this particular phage is so different from the others.

Sporadically distributed "host" genes – a link to cyanobacterial phage–host ecology and evolution

In contrast to the syntenic, widely distributed sets of genes described above, a number of genes exhibit more sporadic distributions across the cyanophage genomes (Table 3), and these are likely driving niche-differentiation of cyanophage–host systems (Lindell et al., 2004; Coleman et al., 2006). Here we highlight a few of these genes, the putative functions of which can be readily connected to known variables in cyanobacterial and cyanophage ecology.

Phosphorus utilization genes: Phosphorus often limits productivity in oligotrophic marine systems, and cyanophages have been shown to contain the phosphate stress gene, *pstS* (Sullivan et al., 2005), which shuttles phosphate from the outer to the inner membrane in cyanobacteria. Two *Prochlorococcus* T4-like phages isolated from the Sargasso Sea have been shown to encode the gene, while it was not found in two *Synechococcus* T4-like phages from coastal waters (Mann et al., 2005; Sullivan et al., 2005;

Weigele et al., 2007, but also see "note in proof"). This raises the question of whether *pstS* distribution is driven by host strain, source waters, or both. Here we observed that homologs of the *pstS* gene were found in 9 of the 16 cyanophages (Table 3). While the 9 phages were isolated on 6 different *Prochlorococcus* and *Synechococcus* host strains, all originated from low nutrient waters, where phosphorus is likely in short supply. Thus it appears that the source waters used for phage isolation are more important than host strain for predicting the presence or absence of *pstS* in the phage genome – a relationship that has been observed in metagenomic analyses of surface ocean samples (Williamson et al., 2008). In addition to the gene itself, we also identified transcriptional regulatory machinery flanking all nine *pstS* genes, including promoters (Fig. 5, Suppl. Fig. 5) and terminators (Fig. 5). No single regulatory solution was apparent across the genomes. Interestingly, two of the phages (S-SM1, S-SM2) contained *phoA*, which encodes an alkaline phosphatase, next to *pstS* (Fig.5). If functional, this could facilitate access to organic phosphorus.

Homologs of *phoH*, a gene which belongs to the phosphate regulon in *E. coli* and encodes a putative ATPase, were found in all 16 cyanophages as well as the marine T4-like vibriophage KVP40 (Miller et al., 2003b). This gene is absent from some other non-T4-like marine cyanophages [e.g., podoviruses P-SSP7 (Sullivan et al., 2005) and P60 (Chen et al. 2002), siphovirus P-SS2 (Sullivan et al., 2009)], but present in other marine phages, i.e., the distant T7-like roseophage SIO1 (Rohwer et al., 2000); thus clear patterns are not evident. We had previously described (Sullivan et al., 2005) such phage-encoded *phoH* genes as apparent parts of a multi-gene family with divergent functions from phospholipid metabolism and RNA modification (COG1702 phoH genes) to fatty acid beta-oxidation (COG1875 phoH genes) (Kazakov et al., 2003); indeed the function of the *phoH* gene, particularly in cyanobacteria, remains unclear. For example, under phosphate stress, the gene has been shown to be upregulated in *E. coli* (Wanner, 1996) and *Corynebacterium glutamicum* (Ishige et al., 2003), down-regulated in *Synechococcus* WH8102 (Tetu et al., 2009), and unaffected in at least two *Prochlorococcus* strains (Martiny et al., 2006). The uniform presence of the gene in the T4-like cyanophages, combined with this mosaic of other patterns of distribution and expression, is intriguing.

Carbon metabolism genes: The distribution of carbon metabolism genes among the cyanophage genomes (Table 3) suggests that many have co-opted critical enzymes to access reducing power from glucose via the pentose phosphate pathway (PPP). All but S-PM2 (Mann et al., 2005) have the transaldolase gene (*talC*), thought to be important in mobilizing stored carbon through the PPP, and observed previously in three T4-like cyanophage genomes (Sullivan et al., 2005; Weigele et al., 2007). These phages also carry the gene that encodes CP12, a cyanobacterial regulatory protein that inhibits several Calvin cycle enzymes, promoting carbon flux through the PPP at night (Tamoi et al., 2005). We recently identified a homolog of CP12 in *Prochlorococcus,* whose identity was strengthened by a diel expression pattern consistent with this function (Zinser et al., 2009). This led to the identification and analysis of *cp12* in these phage genomes, with the diel expression patterns of PPP genes (Zinser et al., 2009) informing their possible role in cyanophages (Thompson et al., in prep.). In addition to carrying *talC* and *cp12*, eight *Synechococcus* cyanophages encode two other pentose phosphate pathway enzymes, of varying sequence conservation (see below), which generate NADPH: *zwf,* a glucose-6-phosphate dehydrogenase, and *gnd,* a 6-phosphogluconate dehydrogenase. The existence of as many as four PPP genes in some phages suggests that this pathway is critical to cyanophage infection. We suggest that this may be due either to increased reducing power stored in carbon substrates or to the production of ribulose-5-phosphate which may alleviate bottlenecks in nucleotide metabolism.

Nitrogen metabolism genes: A well-known cyanobacterial response to nitrogen stress is the degradation of phycobilisomes through the activity of the non-bleaching protein NblA. While the *nblA* gene has been observed in a freshwater cyanophages (Yoshida et al., 2008), this gene has not been found in marine cyanobacteria and has not been observed among marine cyanophage. Here we propose cyanophage involvement in host nitrogen metabolism that likely involves a response to intracellular levels of 2-oxoglutarate (2OG) in the host. Ammonium, the preferred nitrogen source for cyanobacteria, is assimilated through incorporation into a 2OG carbon skeleton. Ammonia limitation thus results in 2OG accumulation in the cell, which serves as an indicator of nitrogen status (Irmler et al., 1997; Forchhammer, 1999; Muro-Pastor et al., 2001). DNA binding of the global nitrogen regulator, NtcA, is 2OG-dependent such that NtcA is inactive when 2OG is limiting and the cell has excess available nitrogen, whereas the opposite is true under nitrogen stress conditions (Schwartz and Forchhammer, 2005).

Three features of the cyanophage genomes suggest that they modulate 2OG levels to stimulate

NtcA activity as needed to promote phage gene expression (Fig. 6). First, all 16 genomes contain numerous NtcA binding sites (1-10 per genome; avg = 4.9), which apparently promote a diversity of both T4 phage and cyanophage functions (Fig. 1). Second, 14 of the 16 genomes contain numerous 2OG-FeII oxygenase superfamily proteins (Table 3). Third, all 16 cyanophages contain at least one and often numerous phytanoyl-CoA-dioxygenases (Suppl. Table 4), enzymes which act on 2OG, in this case as oxidoreductases.

Photosynthesis-related genes: Cyanophage-encoded phycobilin biosynthesis genes have previously been shown to be expressed during infection (*pebS*) and functional in vitro (*pcyA, pebS, ho1*; (Dammeyer et al., 2008). These genes, *pcyA, pebS, ho1*, occur in three, four, and four of the 16 cyanophage genomes, respectively (Table 3). As well, the *cpeT* gene previously observed in S-PM2, S-RSM4 and Syn9 (Mann et al., 2005; Millard et al. 2009; Weigele et al., 2007) is found in 12 of the 16 cyanophage genomes examined here (Table 3). Notably, the *cpeT* gene in marine cyanobacteria is part of a phycoerythrin *cpeESTR* operon, so the role of the cyanophage-encoded copy remains unresolved given the lack of *cpeESR*.

Sporadically distributed among the cyanophage genomes are two electron transport genes, *petE* and PTOX, which encode proteins that commonly co-occur with the carbon metabolism genes (*zwf* and *gnd,* described above) as part of a hypothesized mobile gene cassette (Fig. 7) and likely prevent electrons from backing up and damaging photosynthetic reaction centers. The *petE* gene encodes plastocyanin, and has previously been described in cyanophages (Sullivan et al., 2005; Millard et al. 2009; Weigele et al., 2007). PTOX proteins are normally associated with carotenoid desaturation (Kuntz, 2004), but in cyanophages are hypothesized to help maintain balanced pools of ATP and NADPH in infected host cells (Millard et al. 2009; Weigele et al., 2007). Consistent with this hypothesis, a marine *Synechococcus* was shown recently to use PTOX-related oxidases to shunt off excess inter-photosystem electrons to oxygen rather than to PSI (Bailey et al., 2008), which would significantly impact ATP / NADPH pools. This alternate electron flow was thought to be particularly important under Fe-limiting conditions when PSI/PSII reaction center ratios drop (Bailey et al., 2008). Consistent with this observation, PTOX genes are abundant in open ocean surface water microbial metagenomes (McDonald and Vanlerberghe, 2005), and are found in many surface water oligotrophic *Prochlorococcus* (AS9601, MIT9301, MIT9215, MIT9312, MED4, NATL1A, NATL2A) and *Synechococcus* (BL101, WH8102, CC9902) isolates (data not shown), although lacking in their less Fe-limited counterparts from deeper or coastal waters (e.g., LL *Prochlorococcus* and *SynCC9605*).

Experimentally identified cyanophage structural proteins

To maximize our ability to annotate cyanophage structural proteins, we analyzed the proteome of S-SM1 experimentally, and detected multiple peptides from 41 proteins in the purified S-SM1 virion (Suppl. Table 8, which includes the *Synechococcus* enriched gene T4-GC969 described above). These 41 proteins in S-SM1 and their orthologs in the other 15 cyanophage genomes are designated on Fig. 1 as ORF "underlining", along with the data from two other T4-like phage proteomics projects [S-PM2 (Clokie et al., 2008) and Syn9 (Weigele et al., 2007)]. Notably, these include nine proteins known to be encoded in the S-PM2 genome, but not detected in the virion (Clokie et al., 2008). These nine newly detected proteins encode homologs of seven coliphage T4 structural proteins (gp 4, 5, 14, 21, 25, 48, 53), as well as a two cyanophage core proteins, including a putative citidylyltransferase (T4-GC190) and a hypothetical protein (T4-GC15). We also identified 18 hypothetical proteins which expand the existing dataset of T4-like structural proteins; all of them need structural / functional assignments. We note that 10 virion structural proteins have similar distributions among nine of the cyanophage genomes (Suppl. Table 8); perhaps these proteins are functionally-linked, T4 phage structural components.

Genome evolution in the cyanophages

As discussed above, the "cyanophage core" genes are remarkably syntenic across the 16 cyanophage genomes (Suppl. Fig. 2), suggesting that most of these cyanophage specialization genes are vertically transmitted and part of general T4 phage strategies for infection of ocean cyanobacteria. Twenty-four "core" genes among non-cyanophages were previously inferred to be vertically transmitted and resistant to horizontal gene transfer (Filee et al., 2006; Comeau et al., 2007). It is thought that such genes might be resistant to horizontal gene transfer due to complexity of the T4 protein-protein interactions required for the complex structure (Leiman et al., 2003) and metabolic function (Miller et al., 2003a) of phage T4 and by analogy, the T4-like phages. In contrast, phylogenies of non-core genes in the

T4-like non-cyanophages have conflicting topologies which are interpreted to be due to horizontal gene transfer (Filee et al., 2006). Similarly, our cyanophage core genes are remarkably syntenic, presumably also due to vertical transmission from phage to progeny phage, and the few exceptions to this synteny appear to be due to the activity of mobile genetic elements (Suppl. Fig. 2). Such mobile element activity in T4 phages has been previously observed in coliphage T4 (Miller et al., 2003a), as well as ocean cyanophages ranging from T4-like phages (Zeng et al., 2009) to siphoviruses (Sullivan et al., 2009). Specifically, tRNA genes co-occur with many of these altered non-syntenic regions of the genome (Fig. 1), and may serve as substrates for site-specific recombination by mobile genetic elements (Williams, 2002; Campbell, 2003).

The carbon metabolism genes carried by cyanophages appear particularly influenced by the movement of mobile gene cassettes. For example, *zwf* and *gnd* co-occur in the genomes of eight phages isolated on *Synechococcus* as part of an apparent mobile gene cassette (Fig. 7): five contain paired, full-length, apparently functional gene cassettes in varied genome locations, while three contain variously degraded gene cassettes including remnants of *zwf* genes (Suppl. Fig. 6). The other genes in the apparent mobile cassette include two photosynthetic electron transport genes (*petE* and PTOX, see above), a hypothetical protein (T4-GC119), and an endonuclease, which may at some point have mobilized the cassette as described below. Notably, a ninth genome (*Prochlorococcus* phage P-RSM4) lacks *zwf* and *gnd* entirely, but appears to have remnants of the rest of this mobile cassette (Fig. 7).

The endonucleases in this region are notable as, in phage T4, such genes are known to be part of selfish DNA elements known as intronless homing endonucleases in both coliphages (Belle et al. 2002, Liu et al. 2003) and T4 cyanophages Zeng et al. 2009). It is plausible that such selfish genes might lead to highly recombinogenic regions in the T4 genome as the nuclease errs and yields double strand breaks. Here we observe two forms of endonucleases (Suppl. Fig. 7) – one of which contains sequences with distant homology to this confirmed homing endonuclease (T4-GC228) where only one member (from P-SSM2) contains the catalytic residues identified by Zeng et al. (2009); the second contains sequences that lack any homology to the experimentally determined cyanophage T4 homing endonuclease (T4-GC282). Notably, this endonuclease-flanked mobile gene cassette is located in variable locations in the genomes (Fig. 7). In four of the genomes the cassette appears in the same gp17-gp18 region that Millard et al. (2009) recently described as a hypervariable region. In a fifth genome, S-SM2, the cassette appears near *psbA,* where it is interrupted by a second mobile gene cassette (the hypothetical-T4-GCs cluster described below). The four additional genomes contain degraded forms of this cassette in varied genome locations. Beyond this carbon metabolism cassette, we note that additional carbon metabolism genes, *talC* and *cp12*, occupy variable genome positions ranging from locations in the 5'- or 3'-end of the *psbA* region or near gp5, but are often proximal to tRNAs (Suppl. Fig. 8).

Two other classes of gene cassettes carry signatures of mobility in these genomes. First, a cluster of five hypothetical proteins (T4-GCs 218, 219, 234, 235, 237), often associated with a plasmid stability protein, was found in all but one (S-PM2) of the cyanophages (Suppl. Fig. 9). This cluster was similarly positioned and structured across nine genomes, but varies across the other six genomes. We hypothesize that these proteins are clustered for functional reasons, and that the plasmid stability protein may offer mobility of the gene cassette. Second, large clusters of lipopolysaccharide (LPS) genes are present in the larger cyanophage genomes (Fig. 1) located either near *hli03* (S-SSM7, S-SM2, P-SSM2) and/or near *phoH* (P-SSM2), again proximal to tRNAs. It is not known whether these LPS biosynthesis genes are functional or are simply "stuffer DNA" for headful packaging in these larger genome phages. However, seven LPS genes co-occur in three phages that were isolated two years apart using source waters hundreds of miles distant from each other (T4-GCs 260, 265, 266, 304, 305, 307, 308 all occur in each P-SSM2, S-SM2, S-SSM7). Either a recent transfer event occurred across these three disparate phages, or, perhaps more likely, these LPS genes are functionally linked and represent convergent evolution.

CONCLUSIONS:

With this expanded dataset we have been able to better define the T4-like phage core genome. The challenge now is to examine more closely the non-core genes required for infection of different hosts and environments. Our analysis reinforces the importance, for cyanophage, of carrying genes involved in the light reactions of photosynthesis, the pentose phosphate pathway, and phosphorus acquisition. In addition, we reveal a link to host nitrogen metabolism. Finally, the genome-wide comparison of two phages isolated on the same host from the same sample, offers a first look at *intra*-population genomic

variability that is a critical first step to understanding the biogeography of phage diversity.

NOTE ADDED IN PROOF:

After we completed the analyses of the cyanophage genomes described in this manuscript, another *Synechococcus* phage genome (S-RSM4) became available (Millard et al. 2009). The S-RSM4 genome appears to be a "standard *Synechococcus* T4 phage" as inferred from its genome-wide %G+C (41%) and gene content (contains all 38 T4 core genes, all 25 cyano T4 core genes, all 12 nearly cyano T4 core genes, 21 *Synechococcus* enriched genes, and none of the *Prochlorococcus* enriched genes).

Both S-RSM4 (Millard et al. 2009) and P-RSM5 (this study) were isolated from the oligotrophic Red Sea, and both contain a notable phosphate-related feature. Specifically, P-RSM5, which contains *pstS*, was isolated in September, after months of summer stratification (Lindell & Post 1995, Fuller et al. 2005), which would dramatically reduce nutrient concentrations in surface waters. In contrast, S-RSM4, which lacks *pstS*, was isolated in April before summer stratification (Lindell & Post 1995, Fuller et al. 2005), likely resulting in less stressful nutrient limitation. In fact, cyanobacterial *pstS* expression from these same waters was minimal (Fuller et al. 2005), consistent with a lack of phosphate stress in these waters. We hypothesize, therefore, that the presence/absence of *pstS* in these two phages also reflects the nutrient status of the waters from which they were collected.

MATERIALS AND METHODS:

Phage isolation, purification, DNA extraction and sequencing:

Twelve cyanophages were isolated (Waterbury and Valois, 1993; Sullivan et al., 2003; Sullivan et al., 2008), then concentrated and purified for genomic DNA extraction either by CsCl purification (details in Lindell et al. 2004) or using a Lambda Wizard DNA kit (Promega Corp., Madison, WI) directly on phage lysates. This kit precipitates phage particles using a polyethylene glycol solution, followed by DNA extraction using a diatomaceous earth – based resin (Promega Corp., Madison, WI). Total DNA yields were consistently higher using the Wizard DNA kit than using CsCl-purified particles (1-2 µg from 250 ml lysate vs nanograms from 2 L lysate). Although host DNA contamination was significant (ranged 11.4 - 77.5% of total reads) in the Wizard DNA kit preps due to the less rigorous purification, host reads could be filtered out during phage genome assembly. These methods are described in detail elsewhere (Henn et al. 2010).

Construction and Pyrosequencing Libraries

Pyrosequencing libraries preparations are described in Henn et al. (2010). Briefly, 100 µl of cyanophage genomic DNA (1 ng to 2.2 µg) was sheared using Covaris AFA technology and the following conditions: time = 240 sec, duty cycle = 5, intensity = 5; cycles per burst = 200, and temperature = 3ºC. Post-shearing, the DNA was concentrated and fragments less than 200 bp were removed using AMPure PCR purification beads (Agencourt Bioscience Corporation, Beverly, MA). The DNA shearing profile was determined by running 1 µl of the samples on the Agilent Bioanalyzer 2100 using a DNA 1000 chip (Agilent Technologies, Santa Clara, CA) with the optimal size for library construction was 1.2-1.5 kb fragments. The sheared DNA was then used for pyrosequencing library construction with reagents provided in the GS 20 Library Preparation Kit (454 Life Sciences, Branford CT) according to manufacturer's instructions for fragment end polishing, adaptor ligation, and library immobilization reactions but slightly modified for the clean-up steps, which were performed with the addition of 1.8x AMPure beads.

Genome assembly and annotation

Phage genomes were assembled using the Newbler assembly software package (454 Life Sciences, Branford, CT) with all settings set to default and the '-finish' mode invoked. The '-finish' mode assembles through repetitive regions that form unambiguous paths between contigs, thus some regions that would typically generate an assembly gap were assembled into a contig. Consensus genome sequences reported here represent from 11.9- to 23.8-fold coverage, depending upon the phage, with quality scores better than Q40 for >99.3% of the bases (Henn et al. 2010).

The assembled genomes were annotated in a pseudo-automated pipeline as follows. Open reading frame (ORF) predictions were made using GeneMarkS (Besemer et al., 2001), then manually refined based upon synteny and maximizing ORF size where alternate start sites were present. We next used all predicted ORFs from the 26 T4 phages as BLASTn queries against the genome sequences to

pull out all possible ORFs (e-value cut-off < 1e-5). In this way, we identified a small number of cases (<1%) where the ORF existed in a genome, but had not been predicted by GeneMarkS or manual annotation. Functional annotation to predicted ORFs were assigned using BLASTp (e-value cut-off < 1e-3) against the NCBI non-redundant database (as of April 2009) in combination with gene size and synteny information and HMM profiles for T4-GCs (described below) were HHsearched against the PFAM database. Identification of tRNA genes were done using tRNA-Scan-SE (Lowe and Eddy, 1997). Bacterial sigma-70 promoters and terminators were predicted using BPROM (LDF >2.75, Softberry, Mount Kisco, NY) and TransTermHP (confidence score >80% with an energy score of <-11 and a tail score of <-6; Kingsford et al., 2007), respectively, using default parameters. As well, we specifically searched for known T4 promoters and cyanobacterial nutrient-related promoters as follows. Early T4 phage promoters are sigma-70 promoters that are predicted from the BPROM analysis described above, while to determine T4 late promoters, the known T4 late promoter sequence 5'-TATAAAT-3' (Miller et al., 2003) was used as a query on an initial blastn search (e-value cut-off < 10), over the entire genomes. The resulting sequences were use in a second blastn search (e-value cut-off < 10) to allow for mismatches and obtain further possible promoters. Then only those present in intergenic regions or 10 bp of overlap in the immediate upstream gene were used. Subsequently, known cyanobacterial pho and *ntcA* promoters were identified using consensus sequences for known pho boxes (5'-CTTAN7CTTA-3', (Su et al., 2007)) and using the probabilistic model of *ntcA* binding sites (Su et al., 2005) that was more specifically adapted for use with marine cyanobacteria (5'-GTA-N8-TAC-3'; (Su et al., 2006). In addition to probability scoring cut-offs, all promoters or terminators also were required to be intergenic or within 10 bp of the start/stop of an ORF.

The 12 new cyanophage genome annotations (GU071094-GU071099, GU071101, GU071103, GU071105-GU071106, GU071108, GU075905), and the 4 previously published cyanophage genome annotations (DQ149023, AJ630128, AY940168, AY939844, FM207411) are available at Genbank, while the 10 non-cyanophage genome annotations are available at http://phage.bioc.tulane.edu. Additionally, all 26 T4-like phage genome Genbank accession numbers are available in Table 1, and all 16 new or updated cyanophage genomes are also available as a single project at the CAMERA database (http://web.camera.calit2.net/cameraweb/gwt/org.jcvi.camera.web.gwt.download.ProjectSamplesPage/ProjectSamplesPage.oa?projectSymbol=CAM_PROJ_BroadPhageGenomes).

Whole-genome sequencing of these phages revealed that three previously published gene sequences derived from PCR products from these phages (Sullivan et al. 2006, 2008) were incorrect: *g20* from Syn33 (gene GI:189397306, protein GI:189397307), *g20* from S-SSM7 (gene GI: 189397276, protein GI: 189397277), and *psbA* from S-SSM5 (gene GI:95115381, protein GI:95115382). These previous Genbank accessions for these sequences have been corrected with the sequences from the genomes.

Protein clustering and divergent sequence annotation:

The method for clustering orthologous genes across the 26 T4-like phage genomes was similar to that described previously (Kettler et al., 2007). Briefly, pair-wise orthologous relationships were mapped in all T4-like genomes using reciprocal best BLASTp hit (e-value ≤1e-5) to each other where the sequence alignment length was at least 75% of the protein length of the shorter gene of the two compared. T4 Gene Clusters (T4-GCs) were then built by transitively clustering these orthologs together, where if gene A and B are orthologs and gene B and C are orthologs, then genes A, B, and C are clustered into an orthologous group. To find divergent orthologs missed by the initial BLAST-based approach, we built HMM profiles (Durbin et al., 1998) for the T4-GCs, and then searched singleton T4 genes that were not grouped into any T4-GC against the T4-GC HMM profiles. T4-GC HMM profiles were built by aligning each gene in a T4-GC using MUSCLE version 3.7 (Edgar, 2004) with default parameters and then using hmmbuild from HMMER version 2.3.2 (http://hmmer.janelia.org/) to build the HMM profiles from the resulting alignments. The program hmmsearch also from the HMMER version 2.3.2, was used to search a protein sequence against these in-house T4-GC HMM profiles. Those singletons with significant homology (e-value ≤1e-5) to T4-GC HMMs, were considered for membership in that T4-GC and manually curated to certify membership. A total of 15 single genes were brought into T4-GCs this way.

A multifasta of all ORFs used in this study is provided as a supplementary file which includes in the fasta header the ORF identifier and genome location, T4-GC assignment and functional annotation (Suppl. File 2).

Gene presence/absence network analysis

A presence/absence table of all T4-GCs in the 26 phage genomes was constructed and displayed as a network using the spring-embedded layout option Cytoscape 2.5 (Fig 2) (Cline et al., 2007). This layout option treats the connections (edges) between nodes as springs that repel or attract nodes to each other according to a force function; nodes are positioned to minimize the sum of forces in the network. Nodes in the graph represent the T4-GCs (circles) and the genomes (diamonds), and edges represent the presence of a particular T4-GC in a given genome. Each genome node will therefore have a set of T4-GC nodes connected to it. The resulting network highlights the similarities between genomes based on the presence and absence of gene clusters in each genome.

Virion structural proteomics

Structural proteomic experiments were conducted as described previously (Sullivan et al., 2009). Briefly, the samples were incubated in a denaturing solution of 8 M Urea/1% SDS/100 mM ammonium bicarbonate/10 mM DTT pH 8.5 at 37ºC for 1 hour. Next, the samples were alkylated for one hour by the addition of iodoacetamide to a final concentration of 40 mM and then quenched with 2 M DTT. Following the addition of 4X LDS loading buffer (Invitrogen), each sample was centrifuged at 14,000 rpm for 5 minutes at room temperature, and each sample was fractionated on a NuPAGE 10% Bis-Tris 10 lane gel (Invitrogen) for 2.5 hours at 125 volts, 50 mA and 8 W. Gels were shrunk overnight by the addition of 50% ethanol and 7% acetic acid, and then allowed to swell for 1 hour by the addition of deionized water. Gels were stained with SimplyBlue Safe Stain (Invitrogen) for 2-4 hours, imaged, and sliced horizontally into fragments of equal size based on the molecular weight markers.

In-gel digestion was performed after destaining and rinsing the gel sections with two washes of 50% ethanol and 7% acetic acid, followed by two alternating washes with 50 mM ammonium bicarbonate and acetonitrile. After removal of the last acetonitrile wash, 100 µL of sequencing grade trypsin (Promega) was added to each gel slice at a concentration of 6.6 ng/µL in 50 mM ammonium bicarbonate/10% acetonitrile. The gel slices were allowed to swell for 30 minutes on ice, after which the tubes were incubated at 37ºC for 24 hours. Peptides were extracted with one wash of 100 µL of 50 mM ammonium bicarbonate/10% acetonitrile and one wash of 100 µL of 50% acetonitrile/0.1% formic acid. The extracts were pooled and frozen at -80ºC, lyophilized to dryness and redissolved in 40 µL of 5% acetonitrile, 0.1% formic acid.

Samples were then loaded into a 96-well plate (AbGene) for mass spectrometry analysis on a Thermo Fisher Scientific LTQ-FT. For each run, 10 µL of each reconstituted sample was injected with a Famos Autosampler, and the separation was performed on a 75 mM x 20 cm column packed with C_{18} Magic media (Michrom Biosciences) running at 250 nL/min provided from a Surveyor MS pump with a flow splitter with a gradient of 5-60% water, 0.1% formic acid, acetonitrile 0.1% formic acid over the course of 120 minutes (150 min total run). Between each set of samples, standards from a mixture of 5 angiotensin peptides (Michrom Biosciences) were run for 2.5 hours to ascertain column performance and observe any potential carryover that might have occurred. The LTQ-FT was run in a top five configuration with one MS 200K resolution full scan and five MS/MS scans. Dynamic exclusion was set to 1 with a limit of 180 seconds with early expiration set to 2 full scans.

Peptide identifications were made using SEQUEST (ThermoFisher Scientific) through the Bioworks Browser 3.3. The data was searched with a 10 ppm window on the MS precursor with 0.5 Dalton on the fragment ions with no enzyme specificity. A reverse database strategy (Elias and Gygi, 2007) was employed with a six frame translation of the genomic sequence reversed and concatenated with the forward sequences supplemented with common contaminates and filtered to obtain a false discovery rate of less than or equal to 1%. Peptides passing the filters were mapped back onto the genome and compared to predicted open reading frames.

ACKNOWLEDGEMENTS:

Sequencing of the new phage genomes presented here was supported by the Gordon and Betty Moore Foundation MMI Marine Phage, Virus, and Virome Sequencing Initiative through a grant to MRH. MBS, KHH, MLC, ASD, SEK, LRT, RF, MO, SWC were supported in part by grants to SWC from the Gordon and Betty Moore Foundation, NSF, DOE-GTL, MIT UROP; JCIE was supported by a Fulbright Scholarship, as well as University of Arizona BIO5 and Biosphere 2 funds and NSF (DBI-0850105) to MBS; PW was supported by NIEHS (1-P50-ES012742) and NSF (OCE-0430724). John Waterbury and

Freddy Valois kindly provided 3 cyanophage isolates (Syn1, Syn19, Syn33), and Andy Tolonen and Anton Post kindly collected Red Sea water used for cyanophage isolations.

We thank Brian Binder and the crew of the R/V Endeavor, Dave Karl and the HOT team, Mike Lomas and the BATS team for the sampling opportunities. We also thank Mandy Joye and Matthew Erickson for nitrogen measurements from EN360 cruise, the HOT team and the HOT-DOGS site for nutrient measurements from the HOT179 samples, Jarl Haggerty for assistance automating aspects of genome annotation. As well, the team at Microbes Online (particularly Keith Keller) allowed pre-publication viewing of our cyanophage genomes to leverage the power of their on-line comparative genomics tools, while Virginia Rich, Melissa Duhaime, Li Deng, Qinglu Zeng, and Simon Labrie provided valuable discussion and comments. We thank the Broad Institute Genome Sequencing Platform for their efforts on genome sequencing. Finally, we thank two reviewers and our editor for excellent comments and suggestions on the manuscript, and the editor and Melissa Duhaime for insight into MazG biology.

FIGURE LEGENDS:

Figure 1: Overview of 16 cyanophage genome annotations. Each drawn box represents a predicted open reading frame (ORF) with forward strand ORFs above and reverse strand ORFs below. ORFs are color-coded as per the legend in the figure, while color-coded lines on the genome represent experimentally determined structural proteins (see methods). For spreadsheet version of these data, please see Supplementary File 1.

Figure 2: T4-like gene set relatedness representations. (a) Venn diagram illustrating the hierarchical core gene sets among 26 T4-like genomes. (b) T4-like phage presence/absence gene cluster network. T4 gene clusters (T4-GCs) were used to construct a network of phage genomes and gene clusters found in one or more of the 26 genomes. Genomes are represented as diamonds, with cyanophage genomes colored blue and non-cyanophage colored red. Non-core T4-GCs are represented as a light purple circle, core T4-GCs shared by all genomes are colored dark purple. If a T4-GC is present in a phage genome, an edge (green line) is drawn between that genome and the associated T4-GC. Genomes sharing many T4-GCs are in close spatial proximity to each other in the network. A multifasta file with all ORFs examined in this study is provided to link specific ORFs, T4-GC assignments, and functional annotation (Supplementary File 2).

Figure 3: The core and pan-genomes of the (a) cyanophage and (b) non-cyanophage groups, where the core and pan genomes are represented by square and triangles, respectively. The core and pan-genomes were analyzed for k genomes from cyanophages (n=16) or non-cyanophages (n=10). Each possible variation is shown as a grey point, and the line is drawn through the average. The core genome is defined as genes that are present in the selected k genomes. The pan-genome is the total unique genes found in k genomes. All variations of n choose k: n!/k!(n-k)!.

Figure 4: Whole genome pairwise comparisons across the bounds of the cyano T4 phage genome diversity are examined here. In all three panels, two genomes are compared where lines between the genomes connect homologs, colored ORFs indicate genes that are unique to one genome or the other, and the percent identity of each ORF is plotted in the lower half of each panel. Pairwise genome comparisons are presented for (a) two co-isolated cyanophages, P-HM1 and P-HM2, as well as (b) the three closest non-co-isolated phages, P-RSM4, S-SSM5 and S-SM1, and (c) the three most distant non-co-isolated phages, P-SSM2, S-PM2, Syn9, among the 16 sequenced cyanophage genomes.

Figure 5: Close-up genome representation of the phosphate genes cluster from cyanophages. Genomic features are as described in Fig. 1. To orient the reader to the genome location of the cluster being portrayed, a box is drawn in a reference genome for each or a group of similarly placed phage gene clusters.

Figure 6. Proposed role of 2-oxoglutarate (2OG) during cyanophage infection. (a) In uninfected cyanobacteria, nitrogen limitation causes 2OG to accumulate, leading to 2OG-dependent binding of NtcA to promoters of nitrogen-stress genes, resulting in their expression. (b) Phage infection draws down cellular nitrogen causing N-stress and likely leading to 2OG accumulation. Several cyanophage-encoded enzymes (in bold) suggest that increased 2OG may facilitate phage infection. First, phytanoyl-CoA

dioxygenase converts 2OG to succinate, a major electron donor to respiratory electron transport in cyanobacteria (Cooley and Vermaas, 2001) thus potentially generating energy for the infection process. Second, 2OG-dependent dioxygenase [2OG-Fe(II)] superfamily proteins may function in cyanophage DNA repair (Weigele et al., 2007). Third, cyanophage genomes have multiple NtcA promoters driving genes encoding diverse functions - possibly exploiting the host NtcA-driven N-stress response system.

Figure 7: Close-up genome representation of the carbon metabolic gene cluster from cyanophage genomes. Genomic features are as described in Fig. 1, and genome location orientation is as described for Fig. 5.

Suppl. Fig. 1: Cyanophage genome size plotted as a function of the number of predicted ORFs where original host genera are designated by color.

Suppl. Fig. 2: The genome location of four hierarchical "core" gene sets plotted for 26 T4 phage genomes. Lines connect function-based orthologs across genomes, and are colored as per legend.

Suppl. Fig. 3: Multiple sequence alignment of the T4 phage gp51 baseplate hub catalyst protein from 26 T4 phage genomes. The cyanophage and marine vibriophage copies of gp51 are significantly reduced, missing the first ~200 amino acids relative to the non-cyano non-marine T4 phage copies (the first 140 amino acids of the alignment are not shown). In spite of this size difference, there is marked similarity in the C-terminal region of the protein shown in the alignment.

Suppl. Fig. 4: Maximum likelihood tree of the pyrophosphatase MazG protein. The tree was constructed from 271 aligned amino acids, using PhyML and the JTT model of substitution with gamma-distributed rates empirically estimated from the data. The accession numbers for the sequences used in this analysis are available upon request. Numbers at the nodes represent bootstrap values for 1000 replicates.

Suppl. Fig. 5: Weblogo (http://weblogo.berkeley.edu/) diagrams of the various bioinformatically predicted promoters in the cyanophage genomes.

Suppl. Fig. 6: Multiple sequence alignment of the cyanophage-encoded Zwf proteins identified in varying degrees of preservation across 8 cyanophages. While the sequence conservation is minimal for the three highly degraded copies, their position in the genomes is conserved and remnants of sequence similarity remain along the protein.

Suppl. Fig. 7: Alignment of the endonucleases in T4-GCs 228 and 282. (A) Putative homing endonucleases (T4-GC282) where only the P-SSM2 copy has conserved catalytic residues as compared to the experimentally characterized homing endonuclease present in S-PM2 (S-PM2p177, Zeng et al. 2009). The remaining copies appear to have lost these residues and are likely non-functional, yet are all located at a conserved region suggesting a single evolutionary event of insertion at the 3'-end of gp17 (see upper panel for genome sequence details). (B) Possible endonucleases (T4-GC228) which lack the conserved residues in "A" but nonetheless are highly conserved and proximal to the carbon metabolism genes, suggesting that they may be responsible for genetic shuffling in this region.

Suppl. Fig. 8: Close-up genome representation of the mobile carbon metabolic gene cluster from cyanophage genomes. Genomic features are as described in Fig. 1, and genome location orientation is as described for Fig. 5.

Suppl. Fig. 9: Close-up genome representation of the mobile hypothetical genes cluster from cyanophage genomes. Genomic features are as described in Fig. 1, and genome location orientation is as described for Fig. 5.

SUPPLEMENTARY FILES:
Supplementary File 1: The spreadsheet used to generate the overview of the cyanophage genome annotations that are presented in Figure 1.

Supplementary File 2: Multifasta of all ORFs examined in this study including gene identifiers and genome location, T4-GC assignment and functional annotation.

References:

Angly, F.E., Felts, B., Breitbart, M., Salamon, P., Edwards, R.A., Carlson, C. et al. (2006) The marine viromes of four oceanic regions. *PLoS Biol* **4**: e368.

Arrigo, K.R. (2005) Marine microorganisms and global nutrient cycles. *Nature* **437**: 349-355.

Bailey, S., Melis, A., Mackey, K.R., Cardol, P., Finazzi, G., van Dijken, G. et al. (2008) Alternative photosynthetic electron flow to oxygen in marine *Synechococcus*. *Biochim Biophys Acta* **1777**: 269-276.

Belle, A., Landthaler, M., and Shub, D.A. (2002) intronless homing: site-specific endonuclease SegF of bacteriophage T4 mediates localized marker exclusion analagous to homing endonucleases of group I introns. *Genes and Development* **16**: 351-362.

Besemer, J., Lomsadze, A., and Borodovsky, M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* **29**: 2607-2618.

Breitbart, M., Thompson, L.R., Suttle, C.S., and Sullivan, M.B. (2007) Exploring the vast diversity of marine viruses. *Oceanography* **20**: 353-362.

Breitbart, M., Felts, B., Kelley, S., Mahaffy, J.M., Nulton, J., Salamon, P., and Rohwer, F. (2004) Diversity and population structure of a near-shore marine-sediment viral community. *Proc R Soc Lond B Biol Sci* **271**: 565-574.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segall, A.M., Mead, D. et al. (2002) Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A* **99**: 14250-14255.

Bryan, M.J., Burroughs, N.J., Spence, E.M., Clokie, M.R.J., Mann, N.H., and Bryan, S.J. (2008) Evidence for the intense exchange of *mazG* in marine cyanophages by horizontal gene transfer. *PLoS One* **3**: e2048.

Campbell, A. (2003) Prophage insertion sites. *Res Microbiol* **154**: 277-282.

Chibani-Chennoufi, S., Canchaya, C., Bruttin, A., and Brussow, H. (2004) Comparative genomics of the T4-Like Escherichia coli phage JS98: implications for the evolution of T4 phages. *J. Bacteriology* **186**: 8276-8286.

Cline, M.S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C. et al. (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* **2**: 2366-2382.

Clokie, M.R., and Mann, N.H. (2006) Marine cyanophages and light. *Environ Microbiol* **8**: 2074-2082.

Clokie, M.R., Thalassinos, K., Boulanger, P., Slade, S.E., Stoilova-McPhie, S., Cane, M. et al. (2008) A proteomic approach to the identification of the major virion structural proteins of the marine cyanomyovirus S-PM2. *Microbiology* **154**: 1775-1782.

Clokie, M.R.J., Shan, J., Bailey, S., Jia, Y., and Krisch, H.M. (2006) Transcription of a 'photosynthetic' T4-type phage during infection of a marine cyanobacterium. *Environmental Microbiology* **8**: 827-835.

Coleman, M.L., and Chisholm, S.W. (2007) Code and context: *Prochlorococcus* as a model for cross-scale biology. *Trends Microbiol* **15**: 398-407.

Coleman, M.L., Sullivan, M.B., Martiny, A.C., Steglich, C., Barry, K., Delong, E.F., and Chisholm, S.W. (2006) Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* **311**: 1768-1770.

Comeau, A.M., Bertrand, C., Letarov, A., Tetart, F., and Krisch, H.M. (2007) Modular architecture of the T4 phage superfamily: a conserved core genome and a plastic periphery. *Virology* **362**: 384-396.

Cooley, J.W., and Vermaas, W.F. (2001) Succinate dehydrogenase and other respiratory pathways in thylakoid membranes of Synechocystis sp. strain PCC 6803: capacity comparisons and physiological function. *J Bacteriol* **183**: 4251-4258.

Dammeyer, T., Bagby, S.C., Sullivan, M.B., Chisholm, S.W., and Frankenberg-Dinkel, N. (2008) Efficient phage-mediated pigment biosynthesis in oceanic cyanobacteria. *Curr Biol* **18**: 442-448.

DeLong, E.F., Preston, C.M., Mincer, T., Rich, V., Hallam, S.J., Frigaard, N.U. et al. (2006) Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**: 496-503.

Desplats, C., Dez, C., Tetart, F., Eleaume, H., and Krisch, H.M. (2002) Snapshot of the genome of the pseudo-T-even bacteriophage RB49. *J. Bacteriology* **184**: 2789-2804.

Dufresne, A., Ostrowski, M., Scanlan, D.J., Garczarek, L., Mazard, S., Palenik, B.P. et al. (2008) Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. *Genome Biol* **9**: R90.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998) The theory behind profile HMMs. In *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge, England: Cambridge University Press.

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**: 113.

Elias, J.E., and Gygi, S.P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **4**: 207-214.

Filee, J., Bapteste, E., Susko, E., and Krisch, H.M. (2006) A selective barrier to horizontal gene transfer in the T4-type bacteriophages that has preserved a core genome with the viral replication and structural genes. *Mol Biol Evol* **23**: 1688-1696.

Forchhammer, K. (1999) The PII protein in *Synechococcus* PCC 7942 senses and signals 2-oxoglutarate under ATP-replete conditions. In *The Photosynthetic Prokaryotes*. Peschek, G.A., Loeffelhardt, W., and Schmetterer, G. (eds). New York: Kluwer Academic, pp. 549-553.

Fuhrman, J.A. (1999) Marine viruses and their biogeochemical and ecological effects. *Nature* **399**: 541-548.

Fuller, N.J., West, N.J., Marie, D., Yallop, M., Rivlin, A., Post, A.F., and Scanlan, D.J. (2005) Dynamics of community structure and phosphate status of picocyanobacterial populations in the Gulf of Aqaba, Red Sea. *Limnol. Oceanogr.* **50**: 363-375.

Galperin, M.Y., Moroz, O.V., Wilson, K.S., and Murzin, A.G. (2006) House cleaning, a part of good housekeeping. *Mol Microbiol* **59**: 5-19.

Gross, M., Marianovsky, I., and Glaser, G. (2006) MazG -- a regulator of programmed cell death in Escherichia coli. *Mol Microbiol* **59**: 590-601.

Hambly, E., Tetart, F., Desplats, C., Wilson, W.H., Krisch, H.M., and Mann, N.H. (2001) A conserved genetic module that encodes the major virion components in both the coliphage T4 and the marine cyanophage S-PM2. *Proc Natl Acad Sci U S A* **98**: 11411-11416.

Henn, M., Sullivan, M.B., and al., e. (in prep).

Howard, E.C., Henriksen, J.R., Buchan, A., Reisch, C.R., Burgmann, H., Welsh, R. et al. (2006) Bacterial taxa that limit sulfur flux from the ocean. *Science* **314**: 649-652.

Irmler, A., Sanner, S., Dierks, H., and Forchhammer, K. (1997) Dephosphorylation of the phosphoprotein PII in *Synechococcus* PCC 7942: identification of an ATP and 2-oxoglutarate-regulated phosphatase activity. *Mol Microbiol* **26**: 81-90.

Ishige, T., Krause, M., Bott, M., Wendisch, V.F., and Sahm, H. (2003) The phosphate starvation stimulon of *Corynebacterium glutamicum* determined by DNA microarray analyses. *J Bacteriol* **185**: 4519-4529.

Johnson, Z.I., Zinser, E.R., Coe, A., McNulty, N.P., Woodward, E.M., and Chisholm, S.W. (2006) Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science* **311**: 1737-1740.

Karl, D.M. (2007) Microbial oceanography: paradigms, processes and promise. *Nat Rev Microbiol* **5**: 759-769.

Kazakov, A.E., Vassieva, O., Gelfand, M.S., Osterman, A., and Overbeek, R. (2003) Bioinformatics classification and functional analysis of PhoH homologs. *In Silico Biol* **3**: 3-15.

Kettler, G.C., Martiny, A.C., Huang, K., Zucker, J., Coleman, M.L., Rodrigue, S. et al. (2007) Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet* **3**: e231.

Kingsford, C.L., Ayanbule, K., and Salzberg, S.L. (2007) Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biology* **8**: R22.

Kuntz, M. (2004) Plastid terminal oxidase and its biological significance. *Planta* **218**: 896-899.

Lee, S., Kim, M.H., Kang, B.S., Kim, J.S., Kim, G.H., Kim, Y.G., and Kim, K.J. (2008) Crystal structure of Escherichia coli MazG, the regulator of nutritional stress response. *J Biol Chem* **283**: 15232-15240.

Leiman, P.G., Kanamaru, S., Mesyanzhinov, V.V., Arisaka, F., and Rossmann, M.G. (2003) Structure and morphogenesis of bacteriophage T4. *Cell Mol Life Sci* **60**: 2356-2370.

Li, W.K.W. (1994) Primary production of prochlorophytes, cyanobacteria, and eucaryotic ultraphytoplankton: Measurements from flow cytometric sorting. *Limnology and Oceanography* **39**: 169-175.

Li, W.K.W. (1995) Composition of ultraphytoplankton in the central North Atlantic. *Marine Ecology Progress Series* **122**: 1-8.

Lindell, D., Jaffe, J.D., Johnson, Z.I., Church, G.M., and Chisholm, S.W. (2005) Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* **438**: 86-89.

Lindell, D., Sullivan, M.B., Johnson, Z.I., Tolonen, A.C., Rohwer, F., and Chisholm, S.W. (2004) Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc Natl Acad Sci U S A* **101**: 11013-11018.

Lindell, D., Jaffe, J.D., Coleman, M.L., Futschik, M.E., Axmann, I.M., Rector, T. et al. (2007) Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. *Nature* **449**: 83-86.

Lindell, D., and Post, A.F. (1995) Ultraphytoplankton succession is triggered bydeep winter mixing in the Gulf of Aqaba (Eilat), Red Sea. *Limnol. Oceanogr.* **40**: 1130-1141.

Liu, Q., Belle, A., Shub, D.A., Belfort, M., and Edgell, D.R. (2003) SegG endonuclease promotes marker exclusion and mediates co-converstion from a distant cleavage site. *J. Molecular Biology* **334**: 13-23.

Lowe, T.M., and Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**: 955-964.

Lu, J., Chen, F., and Hodson, R.E. (2001) Distribution, isolation, host specificity, and diversity of cyanophages infecting marine *Synechococcus* spp. in river estuaries. *Applied Environmental Microbiology* **67**: 3285-3290.

Luke, K., Radek, A., Liu, X.P., Campbel, J., Uzan, M., Haselkorn, R., and Kogan, Y. (2002) Microarray analysis of gene expression during bacteriophage T4 infection. *Virology* **299**: 182-191.

Magnusson, L.U., Farewell, A., and Nystrom, T. (2005) ppGpp: a global regulator in Escherichia coli. *Trends Microbiol* **13**: 236-242.

Mann, N.H., Cook, A., Millard, A., Bailey, S., and Clokie, M. (2003) Bacterial photosynthesis genes in a virus. *Nature* **424**: 741.

Mann, N.H., Clokie, M.R., Millard, A., Cook, A., Wilson, W.H., Wheatley, P.J. et al. (2005) The genome of S-PM2, a "photosynthetic" T4-type bacteriophage that infects marine *Synechococcus*. *J. Bacteriology* **187**: 3188-3200.

Marston, M.F., and Sallee, J.L. (2003) Genetic diversity and temporal variation in the cyanophage community infecting marine *Synechococcus* species in Rhode Island's coastal waters. *Applied Environmental Microbiology* **69**: 4639-4647.

Martiny, A.C., Coleman, M.L., and Chisholm, S.W. (2006) Phosphate acquisition genes in *Prochlorococcus* ecotypes: evidence for genome-wide adaptation. *Proc Natl Acad Sci U S A* **103**: 12552-12557.

McDonald, A.E., and Vanlerberghe, G.C. (2005) Alternative oxidase and plastoquinol terminal oxidase in marine prokaryotes of the Sargasso Sea. *Gene* **349**: 15-24.

Millard, A., Clokie, M.R., Shub, D.A., and Mann, N.H. (2004) Genetic organization of the *psbAD* region in phages infecting marine *Synechococcus* strains. *Proc Natl Acad Sci U S A* **101**: 11007-11012.

Millard, A.D., Zwirglmaier, K., Downey, M.J., Mann, N.H., and Scanlan, D.J. (2009) Comparative genomics of marine cyanomyoviruses reveals the widespread occurrence of *Synechococcus* host genes localized to a hyperplastic region: Implications for mechanisms of cyanophage evolution. *Environmental Microbiology*. 11: 2370-2387.

Miller, E.S., Kutter, E., Mosig, G., Arisaka, F., Kunisawa, T., and Ruger, W. (2003a) Bacteriophage T4 genome. *Microb. Mol. Biol. Rev.* **67**: 86-156.

Miller, E.S., Heidelberg, J.F., Eisen, J.A., Nelson, W.C., Durkin, A.S., Ciecko, A. et al. (2003b) Complete genome sequence of the broad-host-range vibriophage KVP40: comparative genomics of a T4-related bacteriophage. *J. Bacteriology* **185**: 5220-5233.

Miller, R.V. (2001) Environmental bacteriophage-host interactions: Factors contribution to natural transduction. *Antonie Van Leeuwenhoek* **79**: 141-147.

Muro-Pastor, M.I., Reyes, J.C., and Florencio, F.J. (2001) Cyanobacteria perceive nitrogen status by sensing intracellular 2-oxoglutarate levels. *J Biol Chem* **276**: 38320-38328.

Nolan, J.M., Petrov, V., Bertrand, C., Krisch, H.M., and Karam, J.D. (2006) Genetic diversity among five T4-like bacteriophages. *Virol J* **3**: 30.

Partensky, F., Hess, W.R., and Vaulot, D. (1999) *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiol. Molc. Biol. Rev.* **63**: 106-127.

Paul, J.H. (1999) Microbial gene transfer: An ecological perspective. *J. Mol. Microbiol. Biotechnol.* **1**: 45-50.

Petrov, V.M., Nolan, J.M., Bertrand, C., Levy, D., Desplats, C., Krisch, H.M., and Karam, J.D. (2006) Plasticity of the gene functions for DNA replication in the T4-like phages. *J Mol Biol* **361**: 46-68.

Rohwer, F., Segall, A., Steward, G., Seguritan, V., Breitbart, M., Wolven, F., and Azam, F. (2000) The complete genomic sequence of the marine phage Roseophage SIO1 shares homology with nonmarine phages. *Limnol. Oceanogr.* **45**: 408-418.

Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S. et al. (2007) The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5**: e77.

Schwartz, R., and Forchhammer, K. (2005) Acclimation of unicellular cyanobacteria to macronutrient deficiency: emergence of a complex network of cellular responses. *Microbiology* **151**: 2503-2514.

Sharon, I., Alperovitch, A., Rohwer, F., Haynes, M., Glaser, F., Atamna-Ismaeel, N. et al. (2009) Photosystem I gene cassettes are present in marine virus genomes. *Nature* **461**: 258-262.

Sharon, I., Tzahor, S., Williamson, S., Shmoish, M., Man-Aharonovich, D., Rusch, D.B. et al. (2007) Viral photosynthetic reaction center genes and transcripts in the marine environment. *ISME J* **1**: 492-501.

Su, Z., Olman, V., and Xu, Y. (2007) Computational prediction of Pho regulons in cyanobacteria. *BMC Genomics* **8**: 156.

Su, Z., Olman, V., Mao, F., and Xu, Y. (2005) Comparative genomics analysis of NtcA regulons in cyanobacteria: regulation of nitrogen assimilation and its coupling to photosynthesis. *Nucleic Acids Res* **33**: 5156-5171.

Su, Z., Mao, F., Dam, P., Wu, H., Olman, V., Paulsen, I.T. et al. (2006) Computational inference and experimental validation of the nitrogen assimilation regulatory network in cyanobacterium *Synechococcus* sp. WH 8102. *Nucleic Acids Res* **34**: 1050-1065.

Sullivan, M.B., Waterbury, J.B., and Chisholm, S.W. (2003) Cyanophages infecting the oceanic cyanobacterium *Prochlorococcus*. *Nature* **424**: 1047-1051.

Sullivan, M.B., Coleman, M., Weigele, P., Rohwer, F., and Chisholm, S.W. (2005) Three *Prochlorococcus* cyanophage genomes: Signature features and ecological interpretations. *PLoS Biology* **3**: e144.

Sullivan, M.B., Lindell, D., Lee, J.A., Thompson, L.R., Bielawski, J.P., and Chisholm, S.W. (2006) Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biology* **4**: e234.

Sullivan, M.B., Krastins, B., Hughes, J.L., Kelly, L., Chase, M., Sarracino, D., and Chisholm, S.W. (2009) The genome and structural proteome of an ocean siphovirus: A new window into the cyanobacterial 'mobilome'. *Environ Microbiol.* **11:** 2935-2951.

Sullivan, M.B., Coleman, M.L., Quinlivan, V., Rosenkrantz, J.R., DeFrancesco, A.S., Tan, G.P. et al. (2008) Portal protein diversity and phage ecology. *Environmental Microbiology* **10**: 2810-2823.

Suttle, C.A. (2005) Viruses in the sea. *Nature* **437**: 356-361.

Suttle, C.A., and Chan, A.M. (1993) Marine cyanophages infecting oceanic and coastal strains of *Synechococcus*: abundance, morphology, cross-infectivity and growth characteristics. *Mar. Ecol. Prog. Ser.* **92**: 99-109.

Tamoi, M., Miyazaki, T., Fukamizo, T., and Shigeoka, S. (2005) The Calvin cycle in cyanobacteria is regulated by CP12 via the NAD(H)/NADP(H) ratio under light/dark conditions. *Plant J* **42**: 504-513.

Tetu, S.G., Brahamsha, B., Johnson, D.A., Tai, V., Phillippy, K., Palenik, B., and Paulsen, I.T. (2009) Microarray analysis of phosphate regulation in the marine cyanobacterium *Synechococcus* sp. WH8102. *ISME J* **3**: 835-849.

Thompson, L.T., and al., e. (in prep.).

Traxler, M.F., Summers, S.M., Nguyen, H.T., Zacharia, V.M., Hightower, G.A., Smith, J.T., and Conway, T. (2008) The global, ppGpp-mediated stringent response to amino acid starvation in Escherichia coli. *Mol Microbiol* **68**: 1128-1148.

Wanner, B.L. (1996) Phosphorus assimilation and control of the phosphate regulon. In *Escherichia coli and Salmonella: Cellular and molecular biology*. Neidhardt, F.C. (ed). Washington DC: ASM Press, pp. 1357-1381.

Waterbury, J.B., and Valois, F.W. (1993) Resistance to co-occurring phages enables marine *Synechococcus* communities to coexist with cyanophage abundant in seawater. *Applied and Environmental Microbiology* **59**: 3393-3399.

Waterbury, J.B., Watson, S.W., Guillard, R.R.L., and Brand, L.E. (1979) Widespread occurrence of a unicellular marine planktonic cyanobacterium. *Nature* **277**: 293-294.

Waterbury, J.B., Watson, S.W., Valois, F.W., and Franks, D.G. (1986) Biological and ecological characterization of the marine unicellular cyanobacterium *Synechococcus*. *Can. Bull. Fish. Aquat. Sci.* **214**: 71-120.

Weigele, P.R., Pope, W.H., Pedulla, M.L., Houtz, J.M., Smith, A.L., Conway, J.F. et al. (2007) Genomic and structural analysis of Syn9, a cyanophage infecting marine *Prochlorococcus* and *Synechococcus*. *Environ Microbiol* **9**: 1675-1695.

Weinbauer, M.G. (2004) Ecology of prokaryotic viruses. *FEMS Microbiol Rev* **28**: 127-181.

Williams, K.P. (2002) Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nuc. Acids Res.* **30**: 866-875.

Williamson, S.J., Rusch, D.B., Yooseph, S., Halpern, A.L., Heidelberg, K.B., Glass, J.I. et al. (2008) The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. *PLoS ONE* **3**: e1456.

Wilson, W.H., Joint, I.R., Carr, N.G., and Mann, N.H. (1993) Isolation and molecular characterization of five marine cyanophages propogated on *Synechococcus* sp. strain WH 7803. *Applied Environmental Microbiology* **59**: 3736-3743.

Wommack, K.E., and Colwell, R.R. (2000) Virioplankton: Viruses in aquatic ecosystems. *Microbiol. Molc. Biol. Rev.* **64**: 69-114.

Yooseph, S., Sutton, G., Rusch, D.B., Halpern, A.L., Williamson, S.J., Remington, K. et al. (2007) The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families. *PLoS Biol* **5**: e16.

Yoshida, T., Nagasaki, K., Takashima, Y., Shirai, Y., Tomaru, Y., Takao, Y. et al. (2008) Ma-LMM01 infecting toxic *Microcystis aeruginosa* illuminates diverse cyanophage genome strategies. *J Bacteriol* **190**: 1762-1772.

Zeidner, G., Bielawski, J.P., Shmoish, M., Scanlan, D.J., Sabehi, G., and Beja, O. (2005) Potential photosynthesis gene recombination between *Prochlorococcus* and *Synechococcus* via viral intermediates. *Environmental Microbiology* **7**: 1505-1513.

Zeng, Q., Bonocora, R.P., and Shub, D.A. (2009) A free-standing homing endonuclease targets an intron insertion site in the *psbA* gene of cyanophages. *Curr Biol* **19**: 218-222.

Zinser, E.R., Lindell, D., Johnson, Z.I., Futschik, M.E., Steglich, C., Coleman, M.L. et al. (2009) Choreography of the transcriptome, photophysiology, and cell cycle of a minimal photoautotroph, *Prochlorococcus*. *PLoS One* **4**: e5135.

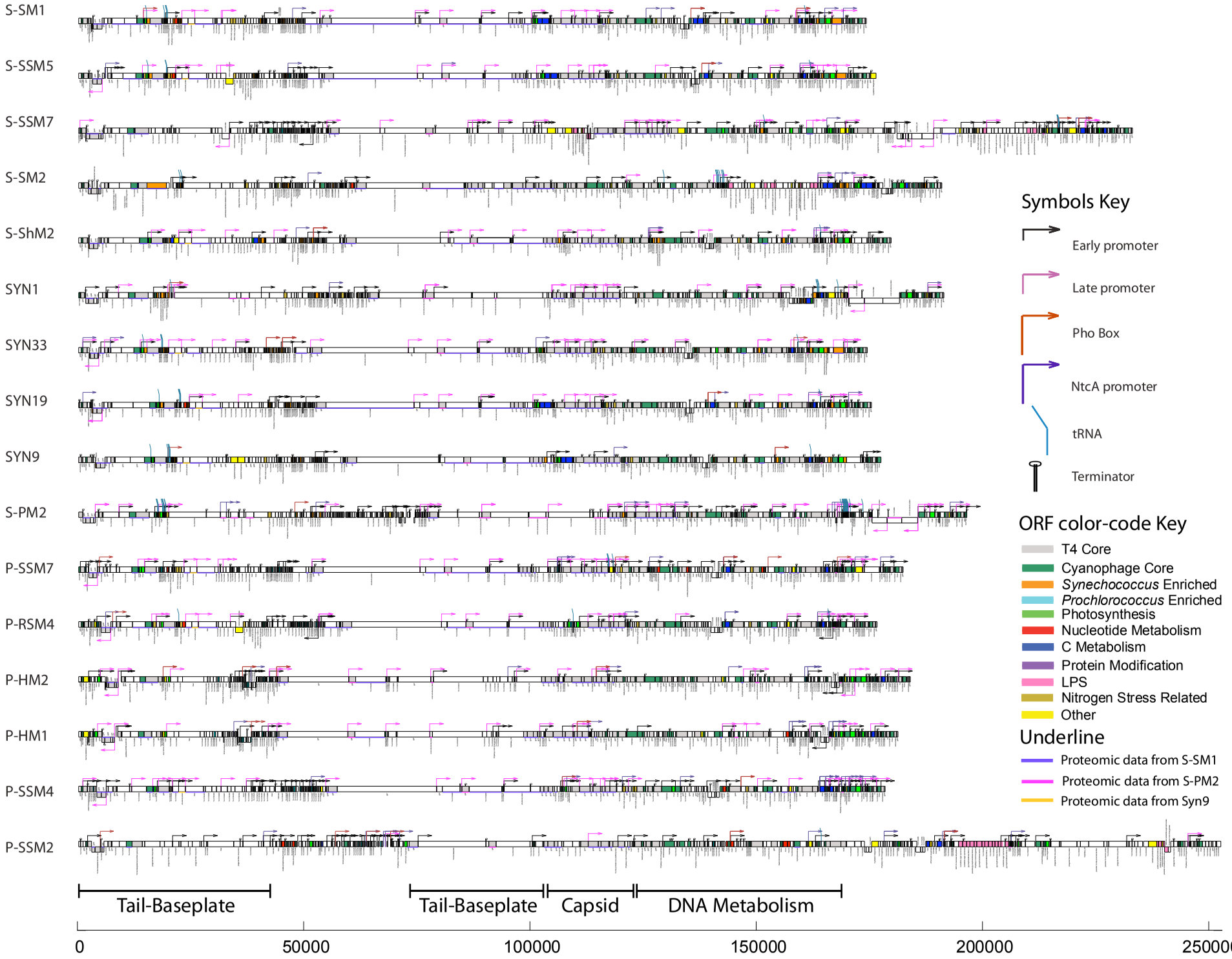

S-SM1
S-SSM5
S-SSM7
S-SM2
S-ShM2
SYN1
SYN33
SYN19
SYN9
S-PM2
P-SSM7
P-RSM4
P-HM2
P-HM1
P-SSM4
P-SSM2
Symbols Key
Early promoter
Late promoter
Pho Box
NtcA promoter
tRNA
Terminator
ORF color-code Key
T4 Core
Cyanophage Core
Synechococcus Enriched
Prochlorococcus Enriched
Photosynthesis
Nucleotide Metabolism
C Metabolism
Protein Modification
LPS
Nitrogen Stress Related
Other
Underline
Proteomic data from S-SM1
Proteomic data from S-PM2
Proteomic data from Syn9
Tail-Baseplate
Tail-Baseplate
Capsid
DNA Metabolism
0
50000
100000
150000
200000
250000

Fig. 2: Overview of gene presence/absence in 26 T4-like phage genomes

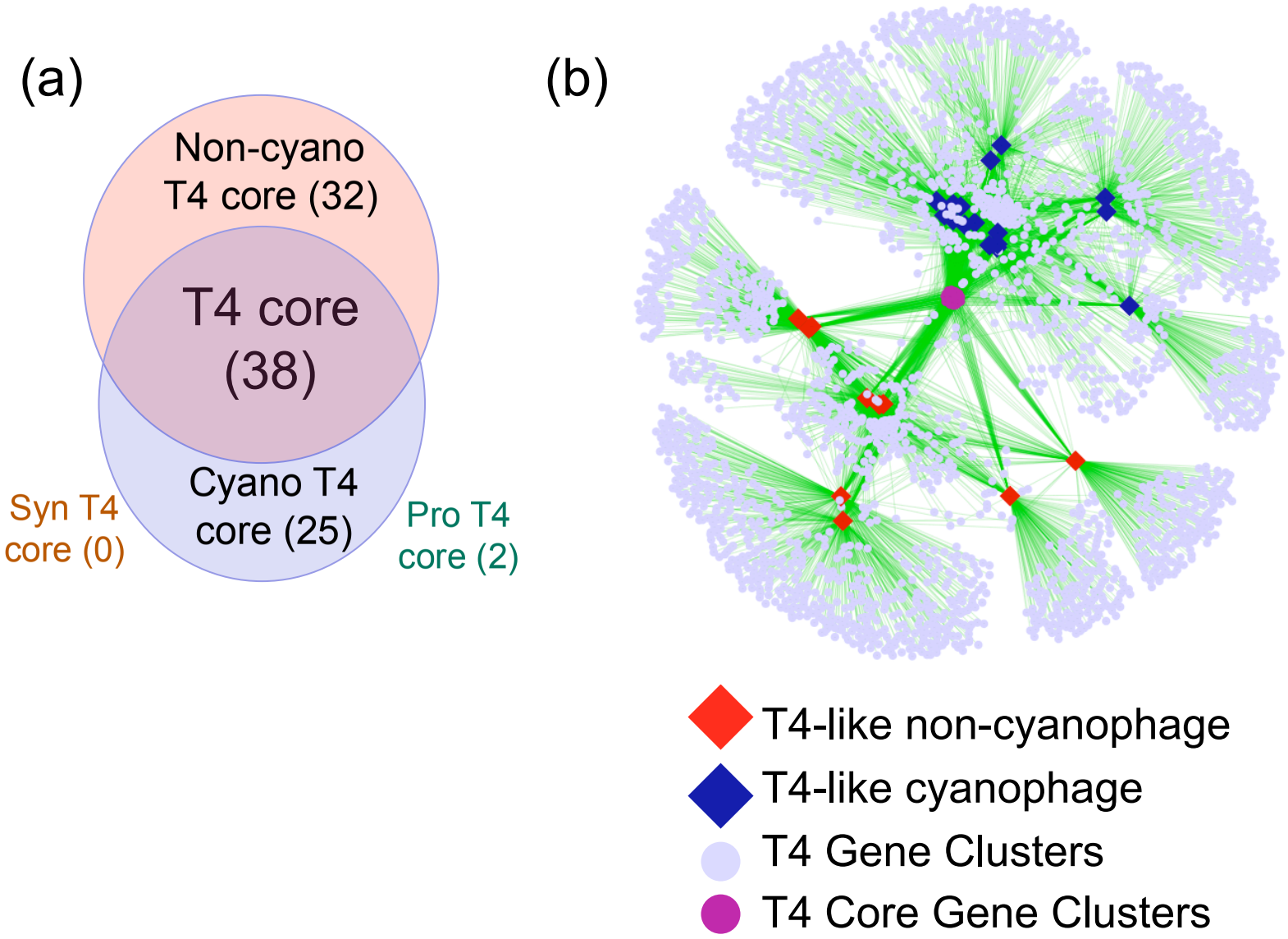


Fig. 3: The Core and Pan genomes of T4-like phages

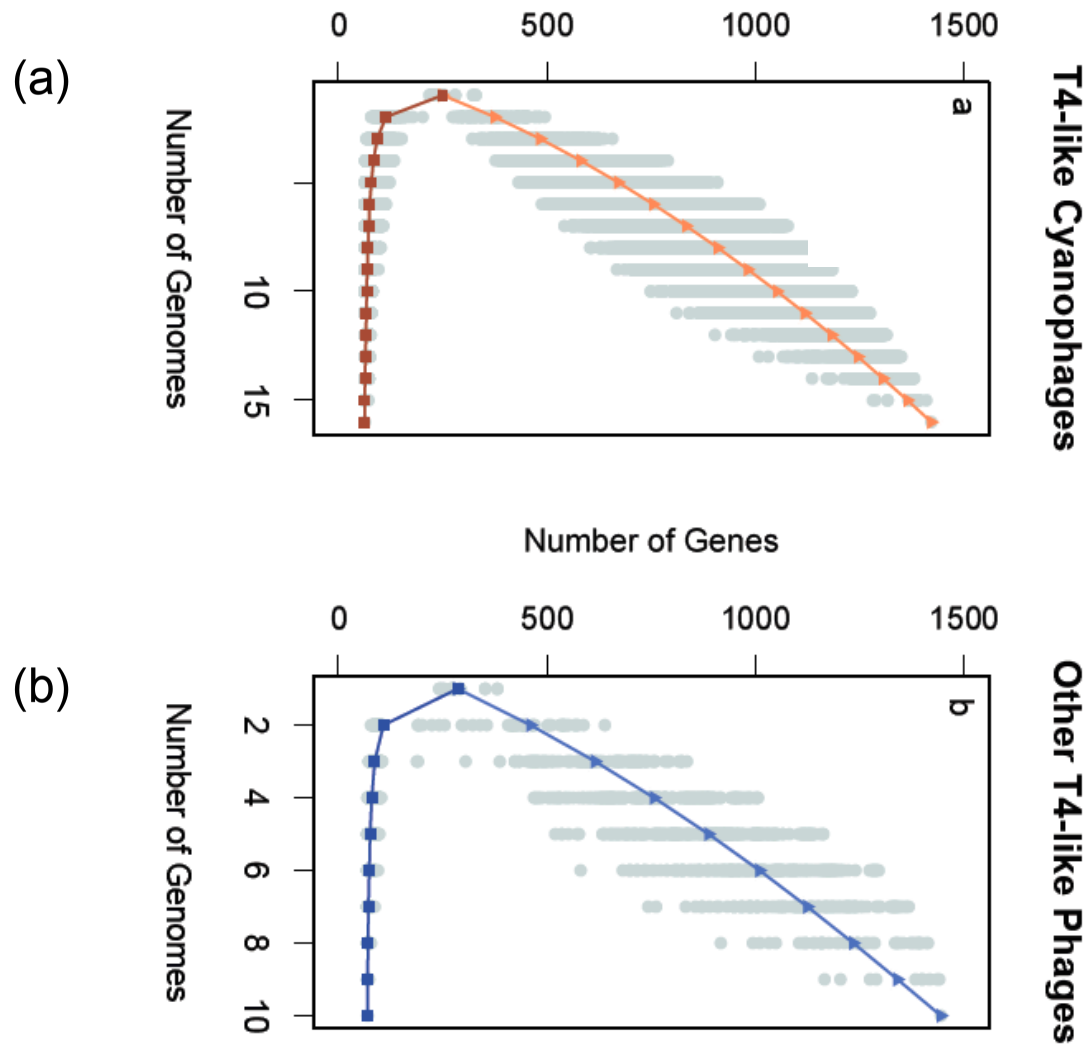
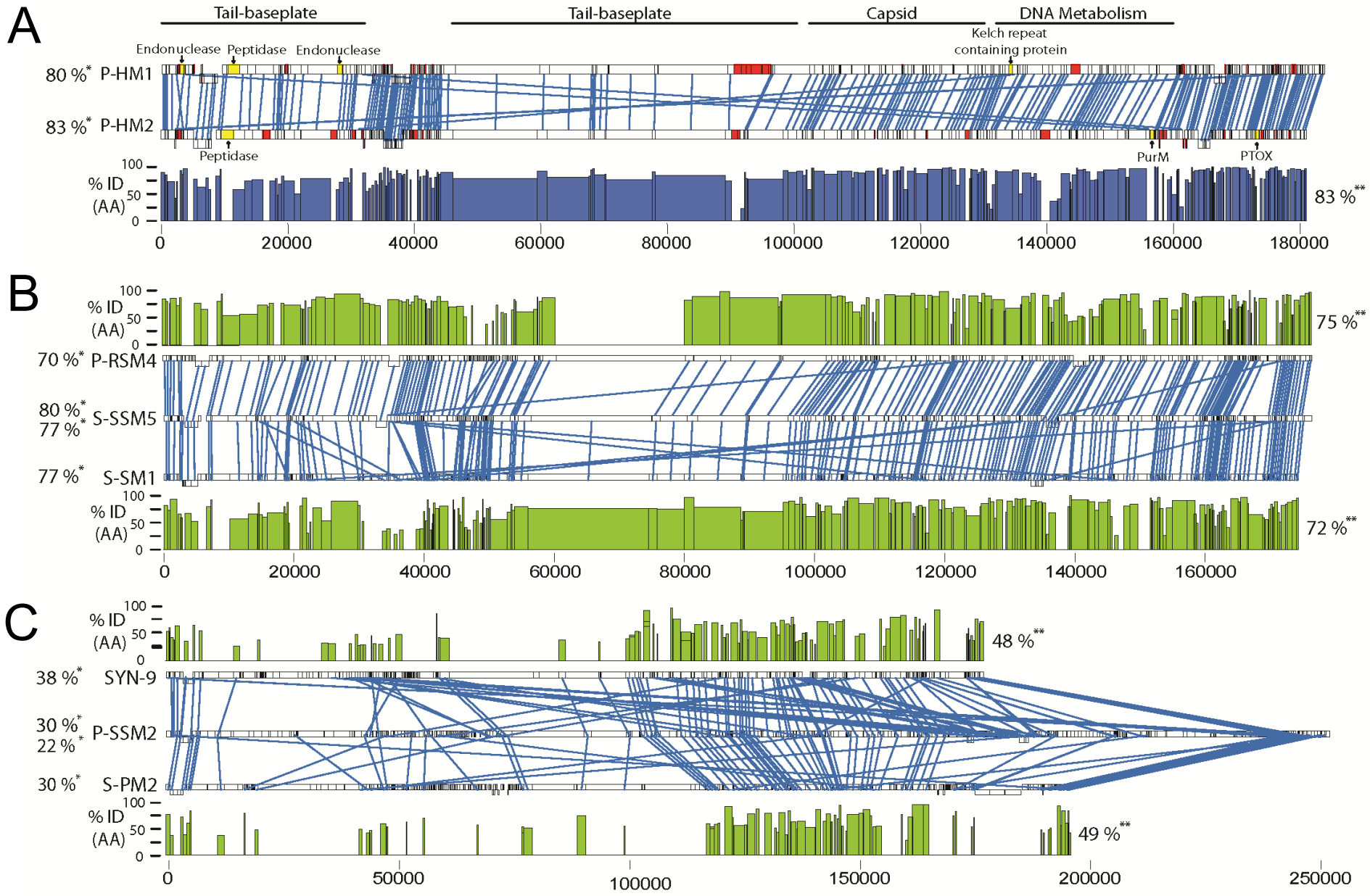


Fig. 4: Whole genome pairwise comparisons



* Percentage of genes shared with the compared genome.

** Average identity across shared proteins.

Fig. 5: Phosphate genes cluster

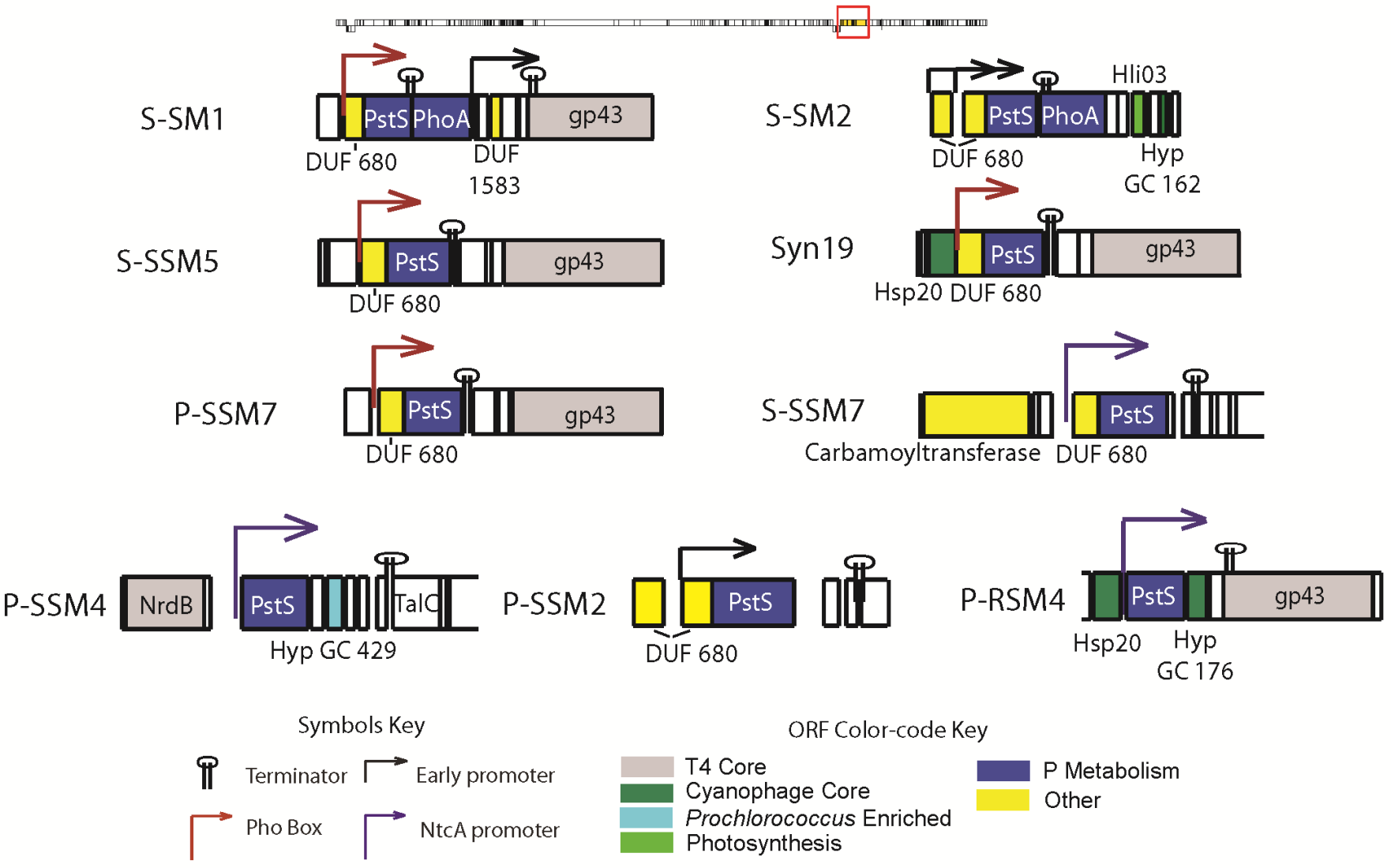


Fig. 6: Nitrogen: A new marine cyanophage link to biogeochemistry

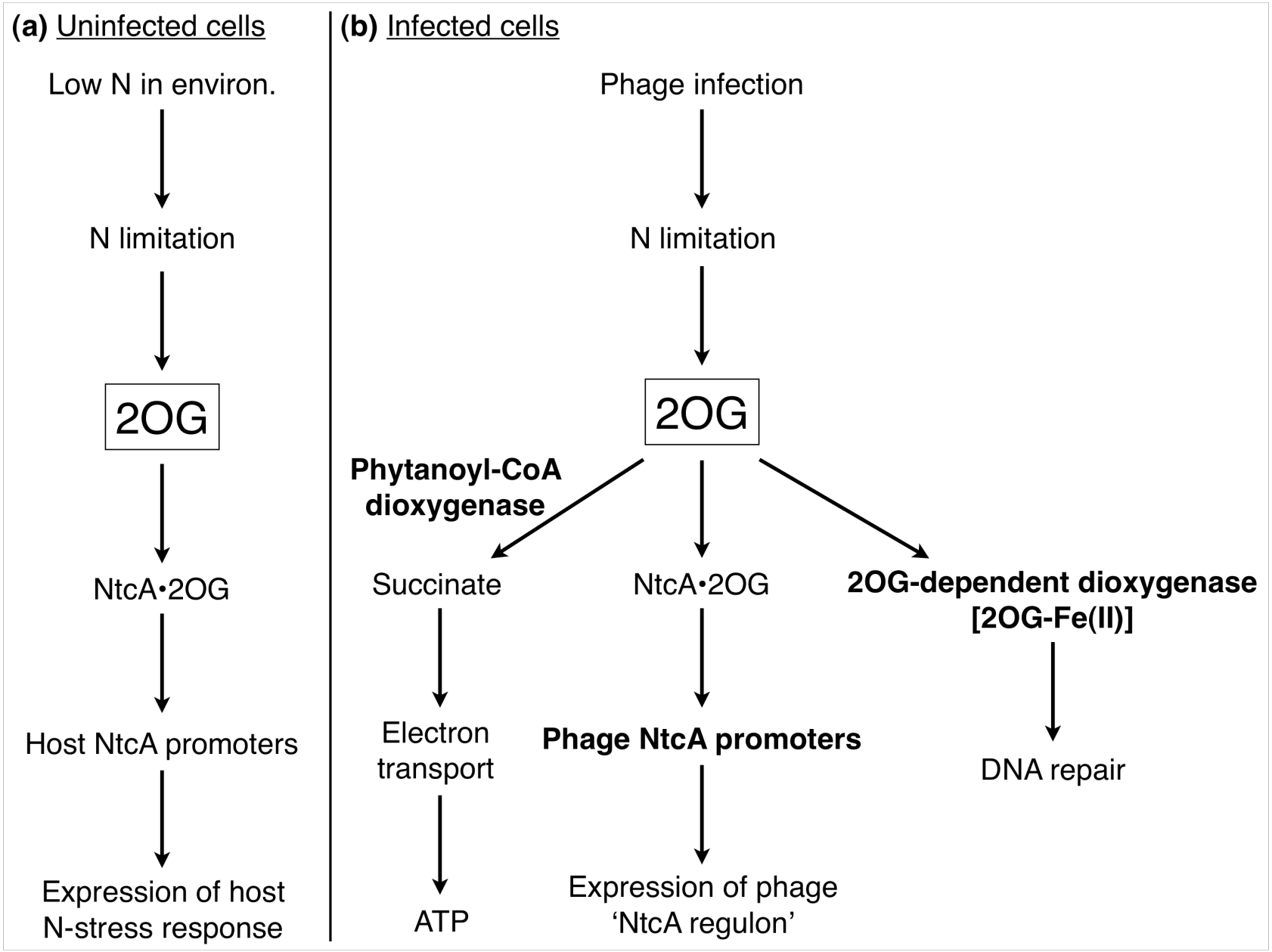
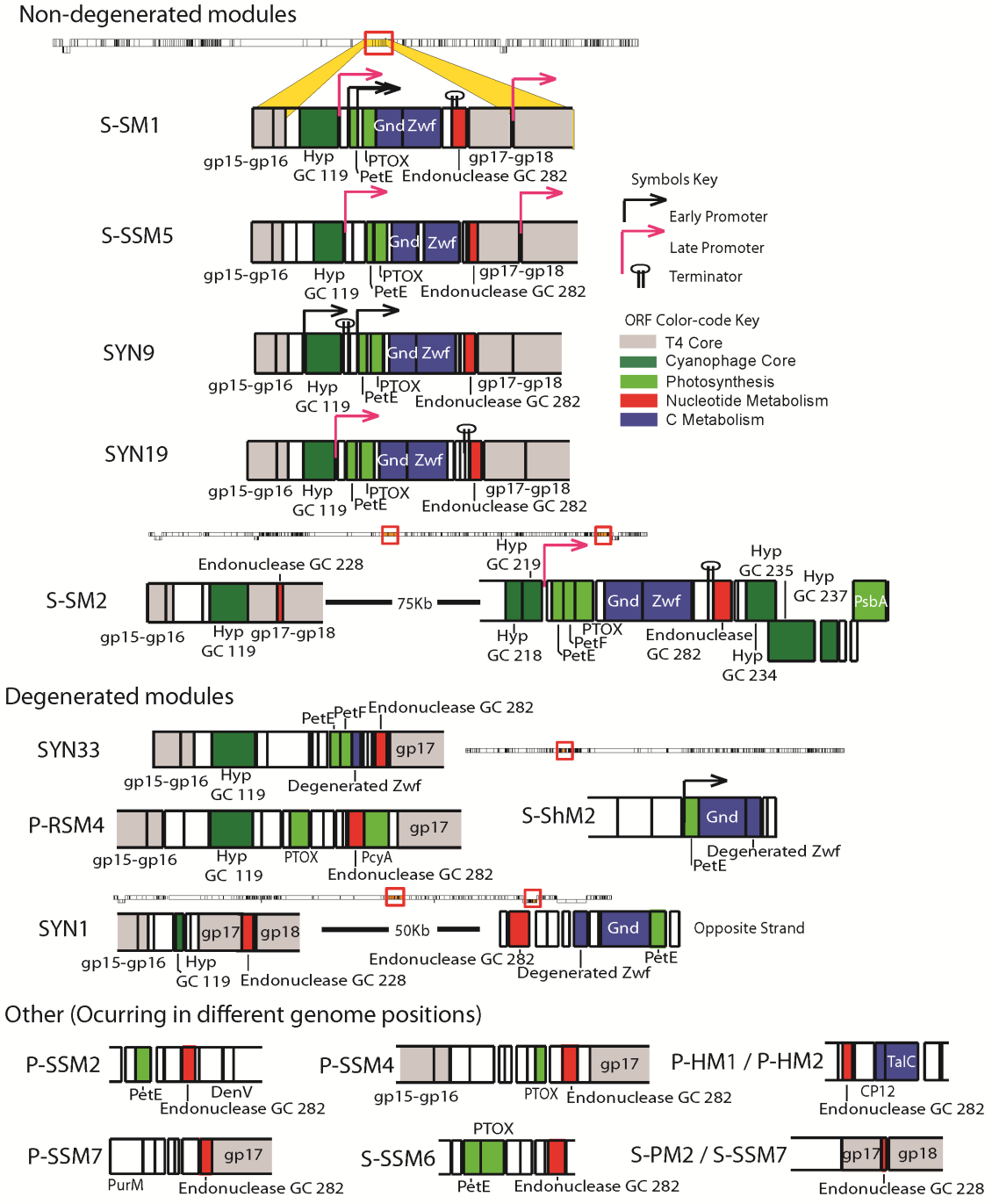
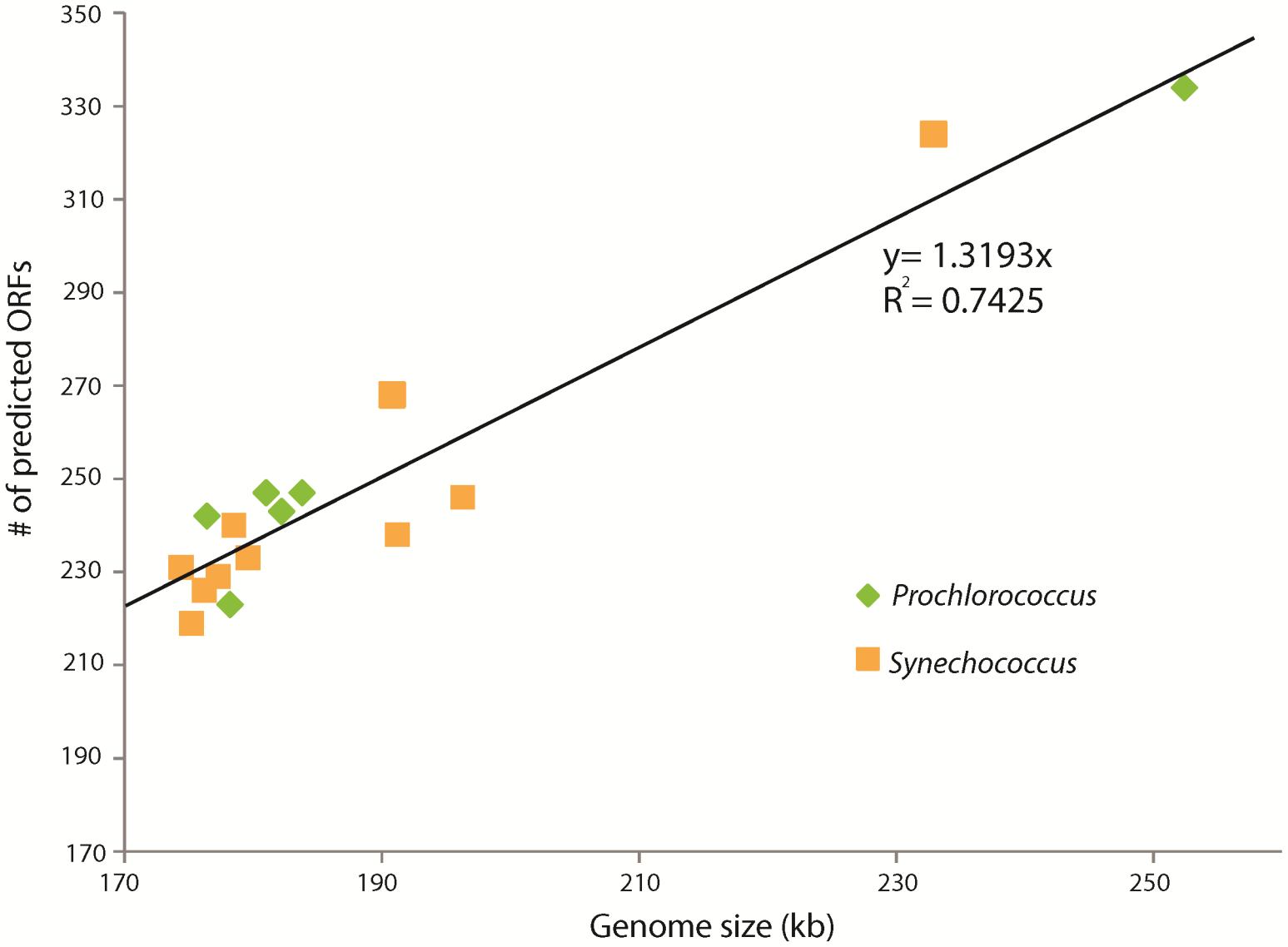


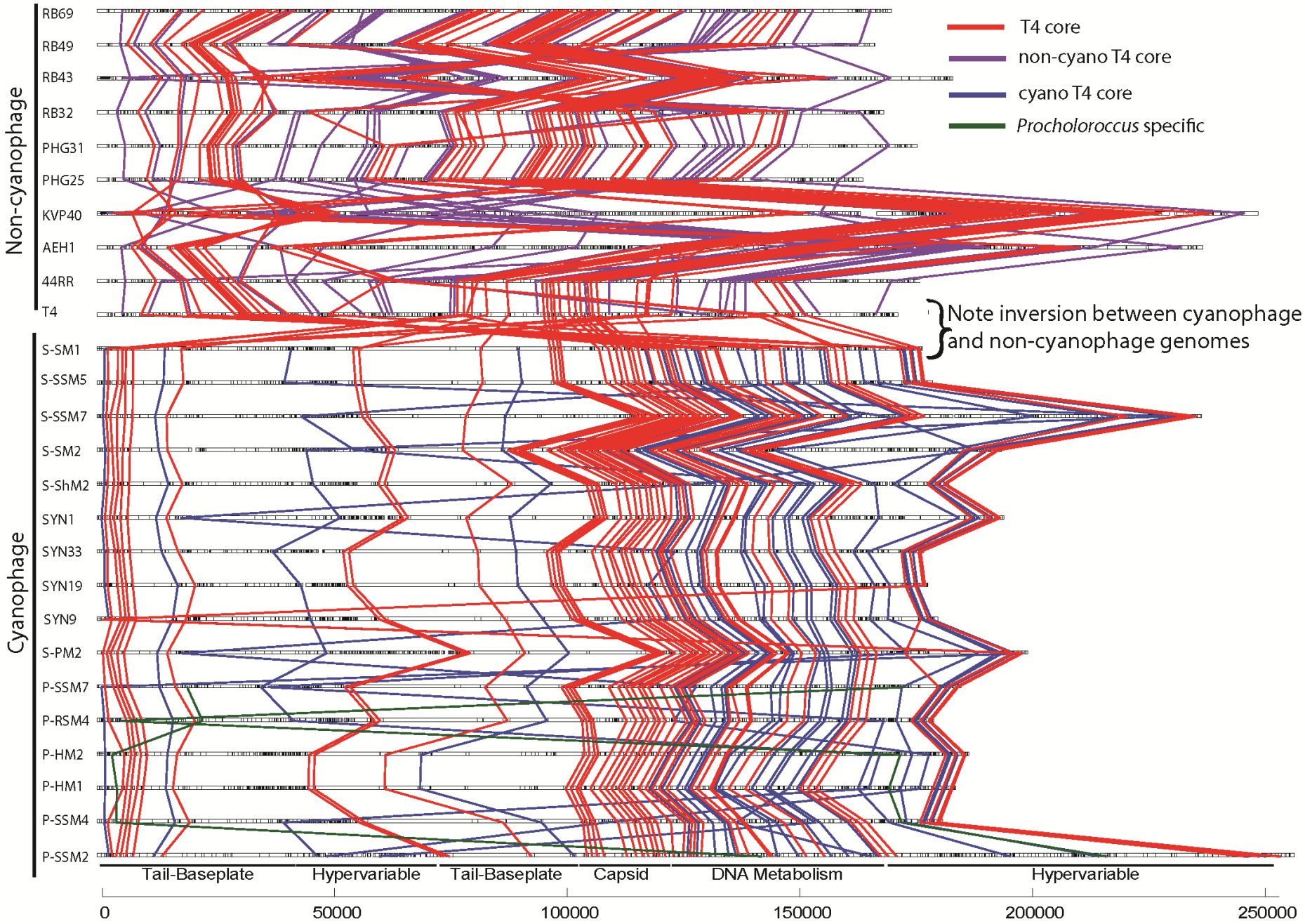
Fig. 7: C-metabolism cluster



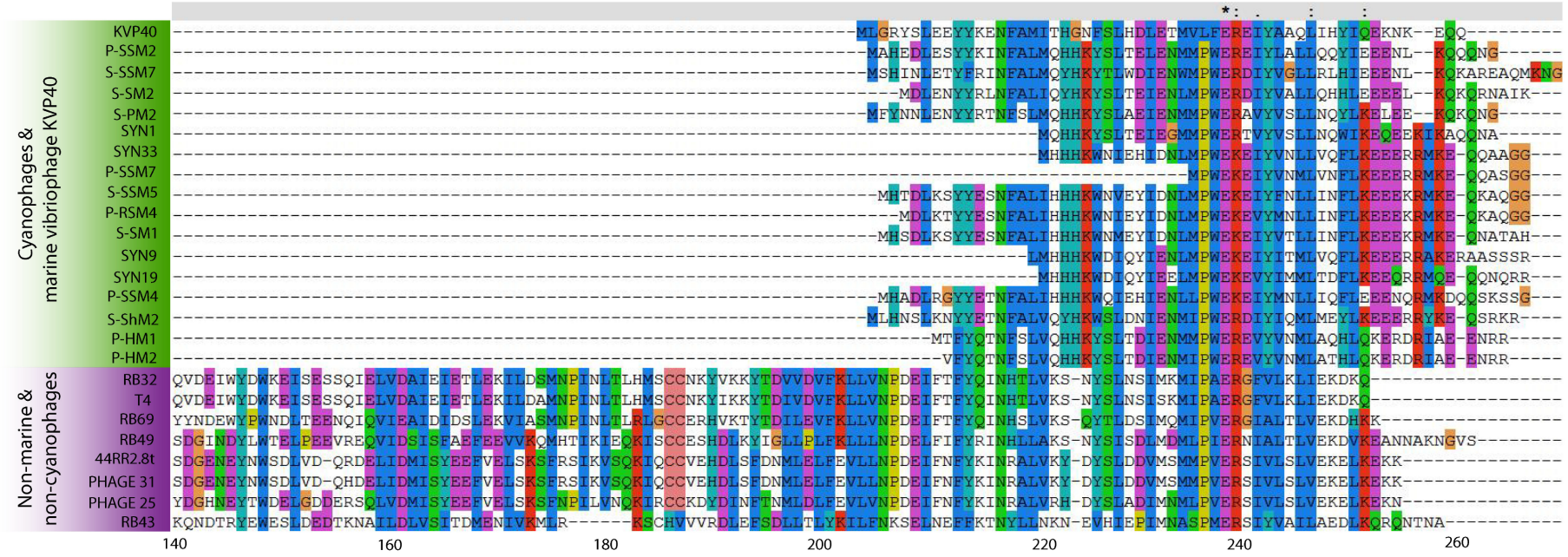
Suppl. Fig. 1: Genome size vs original host / #ORFs



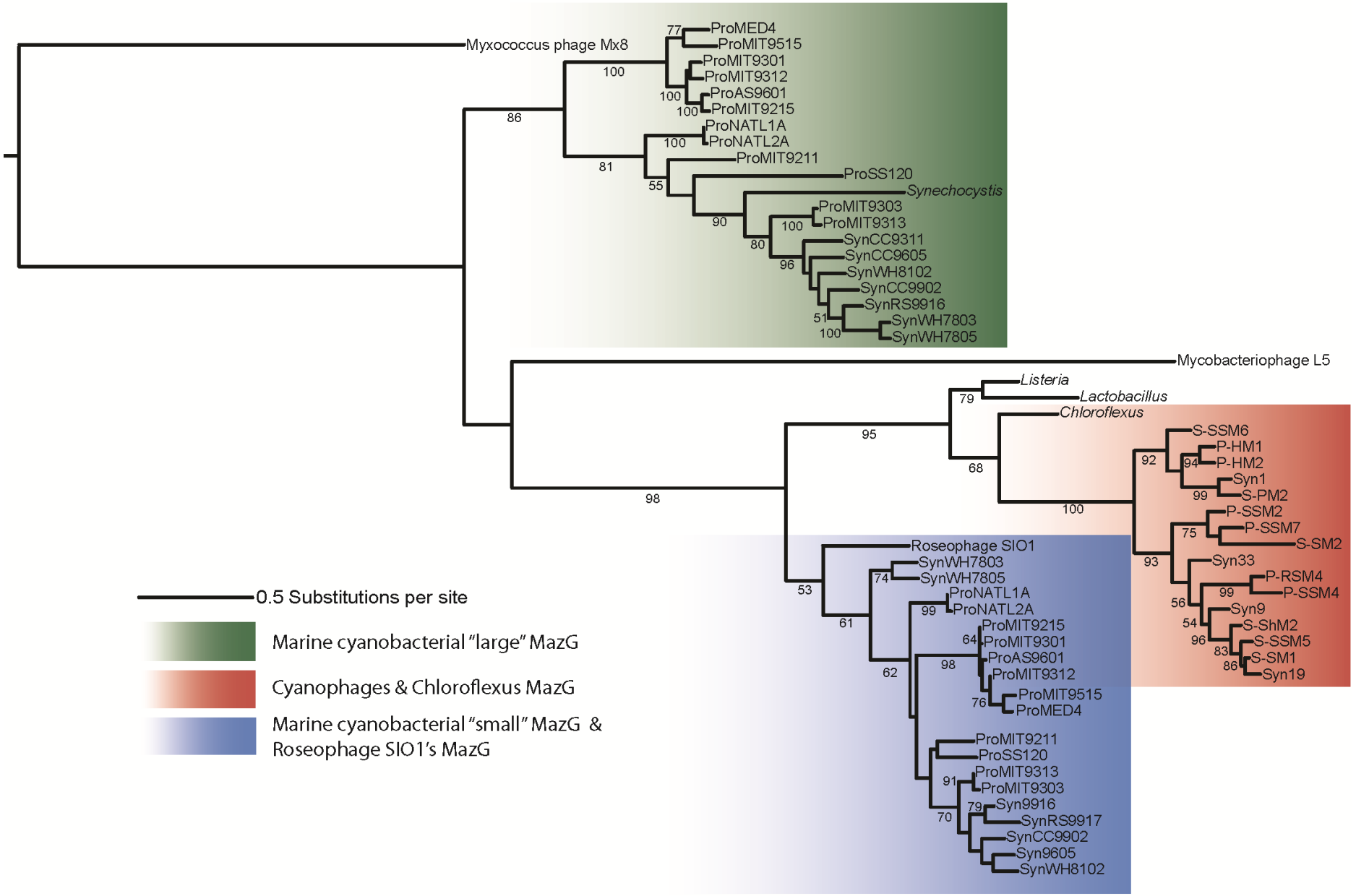
Suppl. Fig. 2: Genome synteny among T4 phages



Suppl. Fig. 3: gp51 alignment



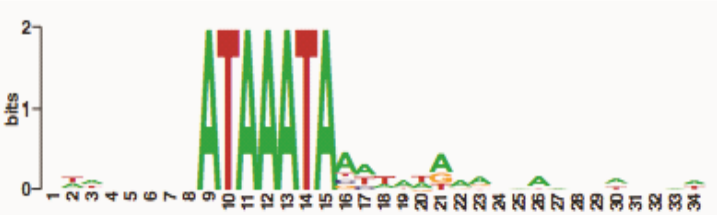
Suppl. Fig. 4: MazG tree



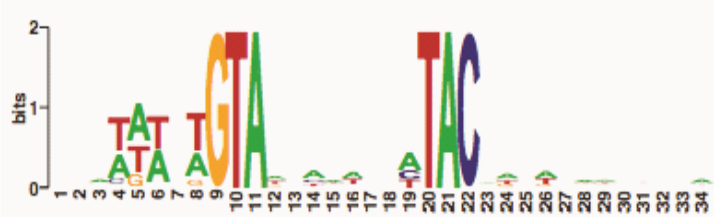
Suppl. Fig. 5: Weblogos of promoters

a) T4-like Early Promoter

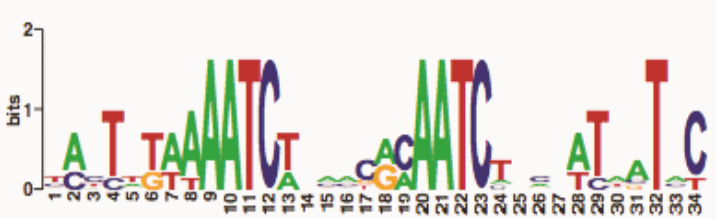
b) T4-like Late Promoter



c) NtcA Promoter



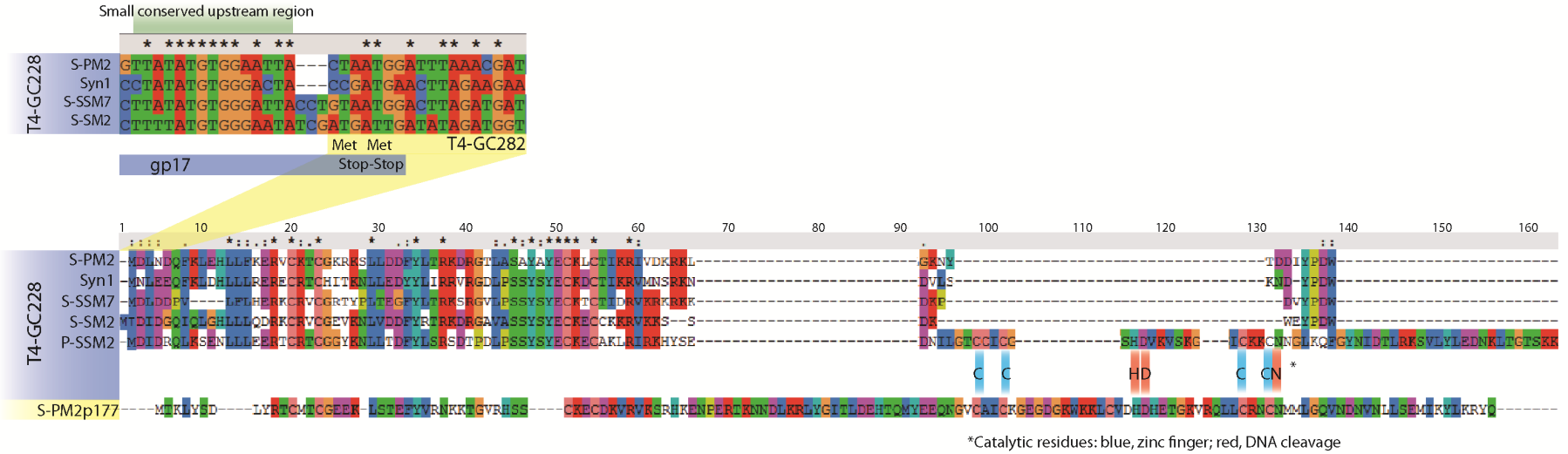
d) Pho Box Promoter



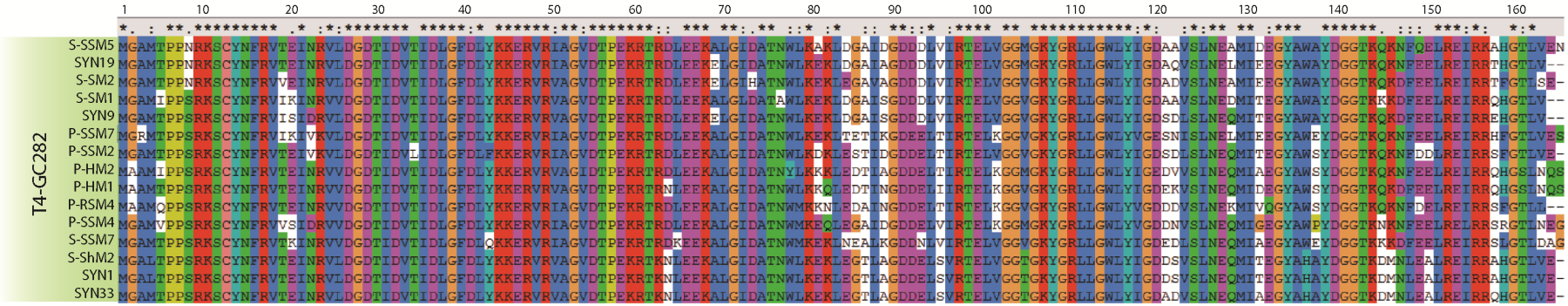
Suppl. Fig. 6: Zwf alignment

Suppl. Fig. 7: Endonuclease alignments

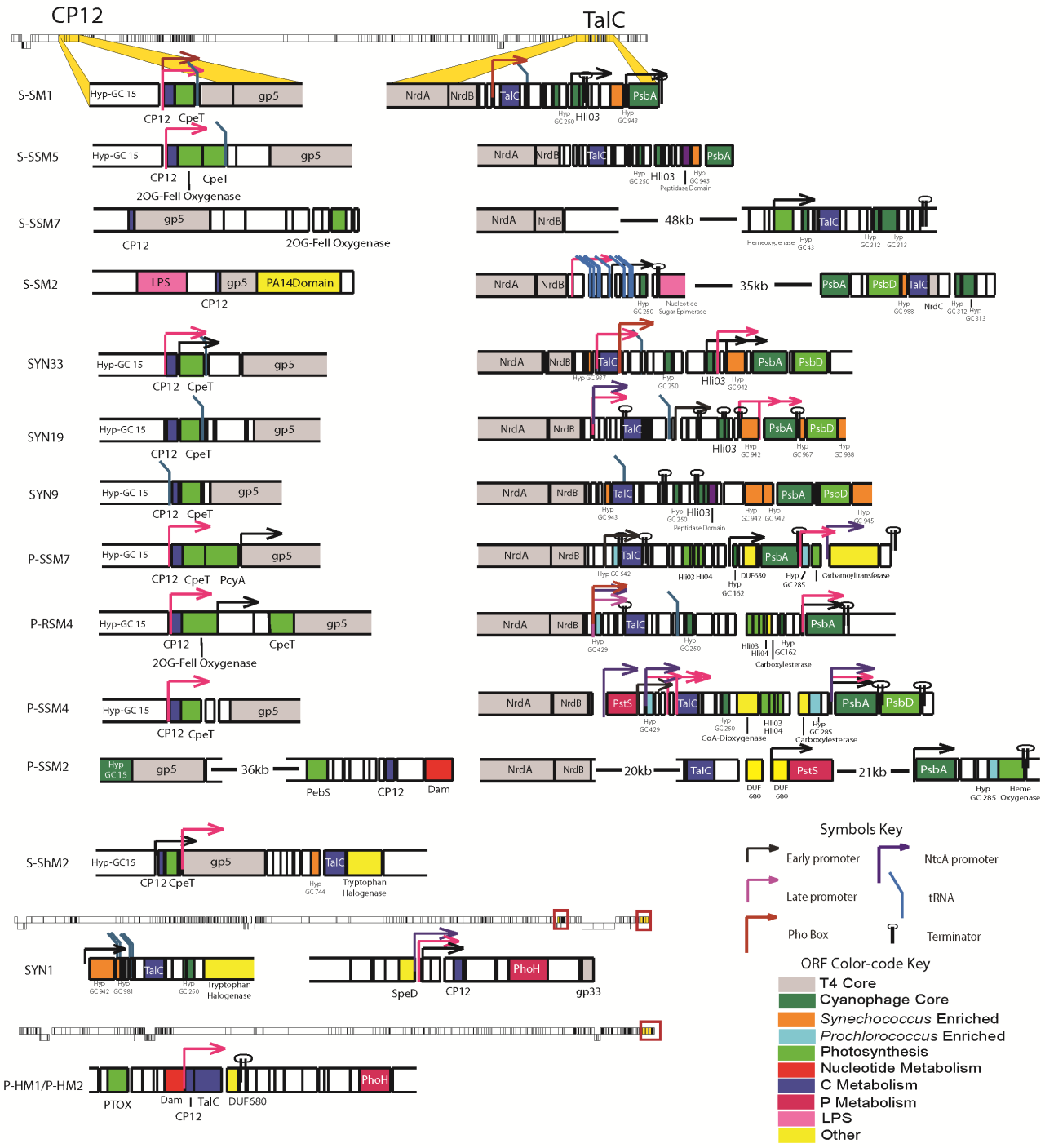
A)



B)



Suppl. Fig. 8: CP12 + talC close-up figure



Suppl. Fig. 9: Mobile Hypothetical Genes Cluster

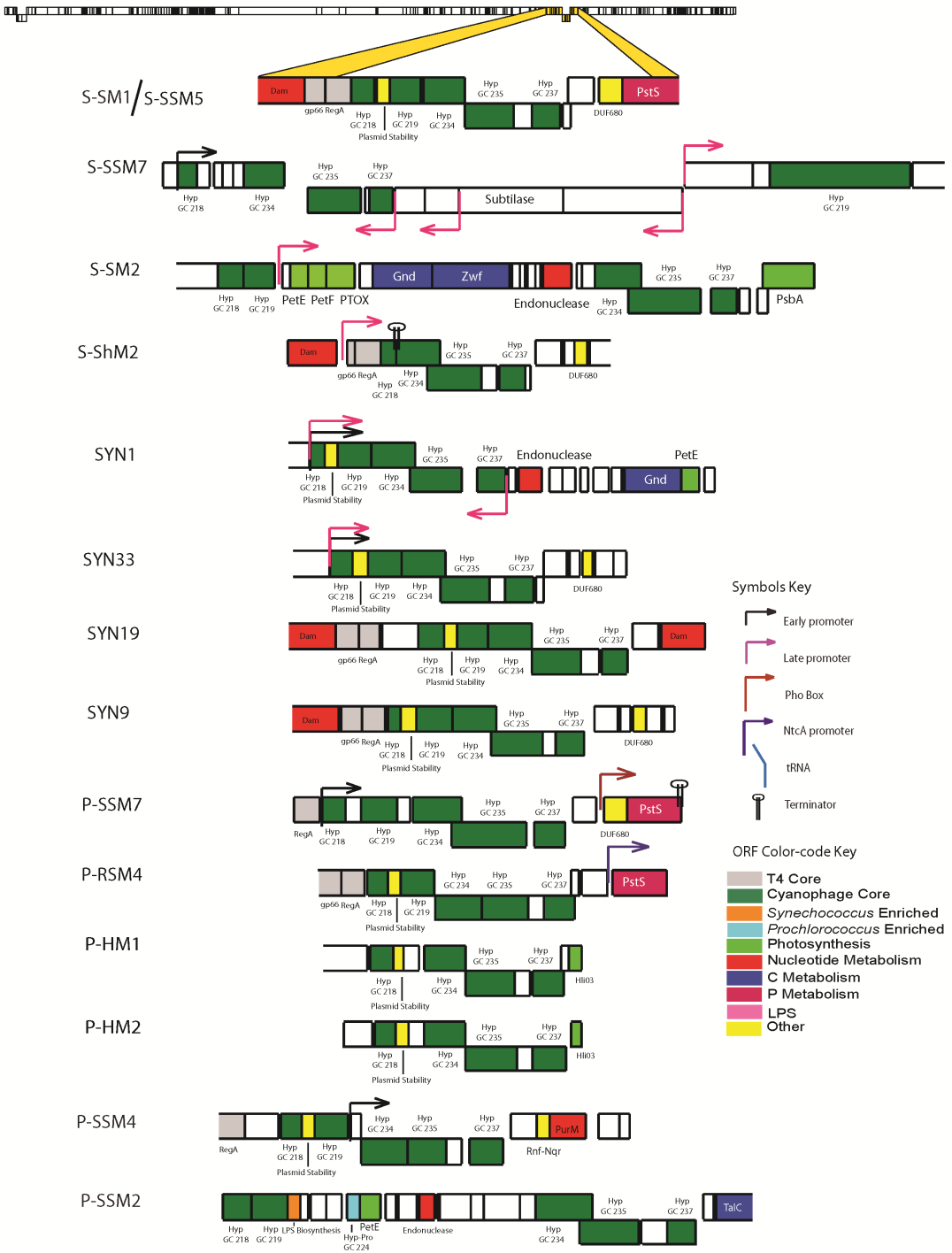


Table 1: General features of the T4-like genomes and isolates.

Published name	Genbank accession #	Original host	Genome Size (kb)	# ORFs	%G+C	Source water description	Date water sampled	# tRNA	Genome publication
Cyanophages									
P-SSM2	AY939844	*Prochlorococcus* NATL1A	252.4	334	35.5%	Atlantic Ocean oligotrophic gyre, 100m	6-Jun-00	1	Sullivan et al. 2005
P-SSM4	AY940168	*Prochlorococcus* NATL2A	178.2	223	36.7%	Atlantic Ocean oligotrophic gyre, 10m	6-Jun-00	0	Sullivan et al. 2005
P-HM1	GU071101	*Prochlorococcus* MED4	181	247	38.0%	Pacific Ocean oligotrophic gyre, 125m	9-Mar-06	0	this study
P-HM2	GU075905	*Prochlorococcus* MED4	183.8	247	38.0%	Pacific Ocean oligotrophic gyre, 125m	9-Mar-06	0	this study
P-RSM4	GU071099	*Prochlorococcus* MIT9303	176.4	242	38.0%	Red Sea, oligotrophic, 130m	13-Sep-00	3	this study
P-SSM7	GU071103	*Prochlorococcus* NATL1A	182.2	243	37.0%	Atlantic Ocean oligotrophic gyre, 120m	Sep-99	4	this study
S-PM2	AJ630128	*Synechococcus* WH7803	196.3	246	37.8%	English Channel, 0m	23-Sep-92	24	Mann et al. 2005
Syn9	DQ149023	*Synechococcus* WH8109	177.3	229	40.50%	Atlantic Ocean coastal (Woods Hole), 0m	Oct-90	6	Weigele et al. 2007
Syn19	GU071106	*Synechococcus* WH8109	175.2	219	41.0%	Atlantic Ocean oligotrophic gyre, 0m	Jul-90	6	this study
Syn33	GU071108	*Synechococcus* WH7803	174.4	231	40.0%	Atlantic Ocean (Gulf Stream), 0m	Jan-95	5	this study
Syn1	GU071105	*Synechococcus* WH8101	191.2	238	41.0%	Atlantic Ocean coastal (Woods Hole), 0m	Aug-90	6	this study
S-ShM2	GU071096	*Synechococcus* WH8102	179.6	233	41.0%	Atlantic Ocean coastal (continental shelf), 0m	16-Sep-01	1	this study
S-SM2	GU071095	*Synechococcus* WH8017	190.8	268	40.0%	15m	17-Sep-01	10	this study
S-SSM7	GU071098	*Synechococcus* WH8109	232.9	324	39.0%	Atlantic Ocean oligotrophic gyre, 70m or 95m	22-Sep-01	5	this study
S-SSM5	GU071097	*Synechococcus* WH8102	176.2	226	40.0%	Atlantic Ocean oligotrophic gyre, 70m	22-Sep-01	4	this study
S-SM1	GU071094	*Synechococcus* WH6501	178.5	240	41.0%	0m	17-Sep-01	6	this study
Non-cyanophages									
T4	AG158101	*E. coli* B	168.9	278	35.3%	likely from sewage see Abedon 2000	N.A.	8	Miller et al. 2003a
RB32	DQ904452	*E. coli*	165.9	270	35.3%	N.A.	N.A.	8	http://phage.bioc.tulane.edu/
RB43	AY967407	*E. coli* B	180.5	292	43.2%	Long Island, NY - sewage	N.A.	1	Nolan et al. 2006
RB49	AY343333	*E. coli* CAJ70	164	274	40.4%	Long Island, NY - sewage	N.A.	0	Nolan et al. 2006
RB69	AY303349	*E. coli* CAJ70	167.6	273	37.7%	Long Island, NY - sewage	N.A.	2	Nolan et al. 2006
KVP40	AY283928	*Vibrio parahaemolyticus*	244.8	381	42.6%	"polluted" coastal seawater off Japan	N.A.	24	Miller et al. 2003b
44RR	AY357531	*Aeromonas salmonicida* 170-68	173.6	252	43.9%	Ontario Canada, Trout pond	N.A.	17	Nolan et al. 2006
Aeh1	AY266303	*Aeromonas hydrophila*	233.2	352	42.8%	Oshkosh, WI - treated sewate effluent	N.A.	23	Nolan et al. 2006
PHG25	DQ529280	*Aeromonas salmonicida* 170-68	161.5	242	41.0%	Eure, France - fish hatchery	N.A.	13	Petrov et al. 2006
PHG31	AY962392	*Aeromonas salmonicida* 95-68	172.9	247	43.9%	Ariege, France - fish hatchery	N.A.	15	Petrov et al. 2006

N.A. = data not available

Table 2: Summary of the 143 "non-core" genes that are enriched in cyanophages (found in >3 genomes), but are absent from non-cyanophages

gene present in # genomes	# of genes	prominent functions (remainder are hypothetical proteins)
4	26	petF, ho1, carbamoyltransferase, pebS, 5 virion structural proteins
5	17	Enase VII, HN, DUF120
6	14	prnA, speD, carboxylesterase, 3 virion structural proteins
7	12	2 virion structural proteins
8	17	purM, 3 virion structural proteins
9	15	pstS, PTOX, 6 virion structural proteins
10	5	petE, 1 virion structural protein
11	5	all hypothetical proteins
12	11	psbD, cpeT, 1 virion structural protein
13	3	denV
14	6	N6A-methylase, helicase, 2OG-FeII oxygenase, 1 virion structural protein
15	12	talC, CP12, DUF680, endonuclease, 1 virion structural protein

Table 3: Summary of cyanobacterial specific sporadically distributed genes among 16 T4-like cyanophages. Presence of the gene occuring in a particular genome is indicated by its size being listed (bp) rather than the lack of the gene indicated by "--". A "|" separates multiple copies of a gene that occur in the same genome.

T4-GC#	FUNCTIONAL ANNOTATION	P-SSM2	P-SSM4	P-HM1	P-HM2	P-RSM4	P-SSM7	S-PM2	Syn9	Syn19	Syn33	Syn1	S-ShM2	S-SM2	S-SSM7	S-SSM5	S-SM1
Photosynthesis																	
440	PsbD = photosystem II D2 protein	--	1077	1077	1077	--	--	1062	1056	1056	1056	1056	1056	1056	--	1056	1056
270+271+274+436	Hli = high-light inducible proteins	114|144|105	135	135	201	219	114	--	--	--	--	--	--	--	--	--	--
404	PTOX = plastoquinol terminal oxidase	--	501	504	--	504	--	--	507	504	--	--	--	504	504	504	504
225	PetE = plastocyanin	345	--	--	--	--	--	--	324	339	324	324	324	324	369	324	351
276	PetF = ferredoxin	294	--	--	--	--	288	--	--	--	291	--	--	294	--	--	--
411	SpeD = S-adenosylmethionine decarboxylase	--	306	--	--	--	--	333	--	--	--	327	--	336	--	342	336
338	CpeT-like protein	--	444	405	405	456	432	528	486	462	459	--	447	--	--	459	453
55	PebS = phycoerythrobilin biosynthesis	702	--	624	624	--	--	--	--	--	--	--	--	--	648	--	--
413	PcyA = phycobilin biosynthesis	--	690	--	--	729	717	--	--	--	--	--	--	--	--	--	--
286+1398	Ho1 = Heme oxygenase	702		693	693	--	--	--	--	--	--	--	--	--	582|165	--	--
615	Hyp. with ferrochetalase domain	--	--	--	--	591	573	609	543	--	--	--	546	732	--	591	531|600
104+240+412+611	2OG-Fe(II) oxygenase superfamily	594|576	567|591|567	--	--	573|711|546|558|600	648|576|579|624	582|717|603	570|606|552	564|597|573	594|594|579	729|609|648	591|600	621	696|618	567|537|558|546|768	582|612|621|573|591
Carbon metabolism																	
920	Gnd = 6-phosphogluconate dehydrogenase	--	--	--	--	--	--	--	1038	1038	--	1038	1023	1038	--	1041	1041
921+1021	Zwf = glucose-6-phosphate dehydrogenase	--	--	--	--	--	--	--	1446	1440	276	303	306	1440	--	1443	1437
337+63	CP12 = carbon metabolic regulator	267	228	213	213	249	246	--	213	285	228	231	228	228	228	228	228
239	TalC = transaldolase	648	654	675	675	678	687	--	702	660	660	648	663	648	654	681	747
Phosphate stress																	
1254	PhoA = alkaline phosphatase	--	--	--	--	--	--	--	--	--	--	--	--	1263	--	--	1356
243	PstS = ABC-type phosphate transport system, substrate binding protein	966	966	--	--	987	966	--	--	981	--	--	--	963	978	981	990
Other functions																	
212	PrnA = tryptophan halogenase	1458	--	--	--	--	--	--	1593|1524	--	--	1437	1116	--	1104|1488	1167	--
425	S-layer domain protein	--	573	--	--	573	633	--	--	--	--	--	--	--	564	573	573
438	carboxylesterase	--	414	375	429	135	--	--	--	--	--	--	--	276	399	--	--
395	HN = hemaglutanin neuraminidase	--	474	474	492	474	--	--	--	--	--	--	--	--	489	--	--
303+721+1350	carbamoyltransferase	1803	--	--	--	1665	1524	--	--	--	--	--	--	--	1803|1536|1575	1665	--
194	tRNA ligase	741	--	--	--	--	798	--	--	--	--	--	--	--	--	--	--

Suppl. Table 1: Detailed features of the T4-like ocean cyanophage isolates

Published name	Original host*	Size (kb)	# ORFs	%G+C	Source water details	Date water sampled	Temp ('C)	Salinity (ppt)	P (umol/kg)	NO3+NO2 (umol/kg)	# tRNA	tRNA's	Ref
Cyanophages													
P-SSM2	ProNATL1A	252.4	330	35.5%	31°48'N, 64°16'W, BATS, 100m	6-Jun-00	19.0	36.7	N.A.	N.A.	1	Asn (AAC);	1
P-SSM4	ProNATL2A	178.2	198	36.7%	31°48'N, 64°16'W, BATS, 10m	6-Jun-00	26.0	36.4	N.A.	N.A.	0	---	1
P-HM1	ProMED4	181	247	38.0%	22°45'N, 158°00'W, Station ALOHA, 125m	9-Mar-06	22.7	35.3	0.05	0.01-0.45	0	---	2
P-HM2	ProMED4	183.8	248	38.0%	22°45'N, 158°00'W, Station ALOHA, 125m	9-Mar-06	22.7	35.3	0.05	0.01-0.45	0	---	2
P-RSM4	ProMIT9303	176.4	246	38.0%	29°28'N, 34°53'E, Red Sea 130m	13-Sep-00	22.0	41.0	N.A.	N.A.	3	Leu (TTA); Arg (AGA); Met (ATG);	2
P-SSM7	ProNATL1A	182.2	241	37.0%	31°48'N, 64°16'W, BATS, 120m	Sep-99	20.3	36.7	N.A.	N.A.	4	Leu (TTA); Arg (AGA); Asn (AAC); Ile (ATA);	2
S-PM2	SynWH7803	196.3	238	37.8%	50°18'N, 4°12'W, English Channel, 0m	23-Sep-92	N.A.	N.A.	N.A.	N.A.	24	Met (ATG) X3; Leu (TTA); Leu (CTA); Arg (AGA); Asn (AAC); Val (GTA); Thr (ACA); Ala (GCA); Gly (GGA); Ile (ATA); Ser (AGC); Ser (TCC); Pro (CCA); Lys (AAA); Tyr (TAC); Asp (GAC); Glu (GAA); Ile (ATC); His (CAC); Gln (CAA); Trp (TGG); Arg(CGT);	3
Syn9	SynWH8109	177.3	235	40.50%	41°31'N, 71°40'W, Woods Hole, 0m	Oct-90	15.0	N.A.	N.A.	N.A.	6	Leu (TTA); Arg (AGA); Asn (AAC); Val (GTA); Thr (ACA); Ala (GCA);	4
Syn19	SynWH8109	175.2	229	41.0%	34°06'N, 61°01'W, Sargasso Sea, 0m	Jul-90	26.5	N.A.	N.A.	N.A.	6	Leu (TTA); Arg (AGA); Asn (AAC); Val (GTA); Thr (ACA); Ala (GCA);	2
Syn33	SynWH7803	174.4	238	40.0%	25°51'N, 79°26'W, Gulf Stream, 0m	Jan-95		N.A.	N.A.	N.A.	5	Leu (TTA); Arg (AGA); Asn (AAC); Val (GTA); Thr (ACA);	2
Syn1	SynWH8101	191.2	224	41.0%	41°31'N, 71°40'W, Woods Hole, 0m	Aug-90	23.0	N.A.	N.A.	N.A.	6	Leu (CTA); Arg (AGA); Asn (AAC); Val (GTA); Thr (ACA); Gly (GGA);	2
S-ShM2	SynWH8102	179.6	231	41.0%	39°60'N, 71°48'W, Atlantic Shelf Waters, 0m	16-Sep-01	20.7	33.4	N.A.	0.043	1	Arg (AGA);	2
S-SM2	SynWH8017	190.8	292	40.0%	38°10'N, 73°09'W, Atlantic Slope Waters, 15m	17-Sep-01	24.0	35.9	N.A.	0.049	10	Leu (TTA); Arg (AGA); Asn (AAC); Val (GTA); Thr (ACA); Ala (GCA); Gly (GGA); Ile (ATA); Ser (TCA); Pro (CCA);	2
S-SSM7	SynWH8109	232.9	324	39.0%	34°24'N, 72°03'W, W Sargasso Sea, 70m or 95m	22-Sep-01	22.0	36.8	N.A.	N.A.	5	Leu (TTA); Arg (AGA); Ile (ATA); Thr (ACA); Gly (GGA);	2
S-SSM5	SynWH8102	176.2	229	40.0%	34°24'N, 72°03'W, W Sargasso Sea, 70m	22-Sep-01	23.7	36.7	N.A.	N.A.	4	Leu (TTA); Arg (AGA); Val (GTA); Thr (ACA);	2
S-SM1	SynWH6501	178.5	239	41.0%	38°10'N, 73°09'W, Atlantic Slope Waters, 0m	17-Sep-01	24.0	35.9	N.A.	0.011	6	Leu (TTA); Arg (AGA); Asn (AAC); Val (GTA); Thr (ACA); Ala (GCA);	2

* original hosts are either genus *Prochlorococcus* indicated by "Pro" or *Synechococcus* indicated by "Syn"

References: 1 = Sullivan et al. 2005, 2 = this study, 3 = Mann et al. 2007, 4 = Weigele et al. 2007

N.A. = data not available

Suppl. Table 2: T4-like phage core genes determined from 16 cyanophages and 10 non-cyanophages***. Numbers listed for each phage represent the size of the genes (bp), with multiple copies separated by a "|". Some T4-GCs were pooled to create a single functional category based upon annotation and genome synteny.

		cyanophages																non-cyanophages									
T4-GC #	GENE DESCRIPTION	P-SSM2	P-SSM4	P-HM1	P-HM2	P-RSM4	P-SSM7	S-PM2	Syn9	Syn19	Syn33	Syn1	S-ShM2	S-SM2	S-SSM7	S-SSM5	S-SM1	T4	44RR	Aeh1	KVP40	PHG25	PHG31	RB32	RB43	RB49	RB69
133	gp3 head-proximal tip of tail tube tail completion + sheath stabilizer protein	534	552	558	546	549	549	510	573	570	549	506	567	576	561	549	549	531	528	570	534	528	528	531	534	591	585
9	gp4 head completion protein	438	426	435	462	426	420	438	420	420	420	342	438	444	480	426	474	453	453	471	456	453	453	453	435	474	450
340+455+16+ 1156+2374	gp5 baseplate hub + tail lysozyme	2259	2310	1764|870	2787|873	2541	2553	2946	2508	2385	2580	2922	2484	1848	3027	2544	2553	1728	1803	1815	1266	1674	1803	1728	1770	1803	1734
106	gp6 baseplate wedge	1944	1989	1851	1851	2028	2031	1809	2031	2025	2031	1809	2133	2112	1875	2028	2019	1983	1884	1956	1959	1884	1884	1983	1854	1905	1971
108+1594	gp8 baseplate wedge	1602	1533	1515	1515	1539	1533	1905	1533	1533	1533	1884	1533	1551	1596	1533	1533	1005	987	987	1032	987	987	1005	993	996	1005
116	gp13 neck protein	846	807	816	816	810	807	831	807	807	807	831	807	846	819	807	807	930	924	921	924	945	924	930	924	933	927
117	gp14 neck protein	1413	927	1329	1329	933	927	879	1173	921	1173	879	1164	1455	2289	927	927	771	759	789	837	762	759	771	747	741	765
118	gp15 proximal tail sheath stabilization	843	1032	1005	1005	786	789	801	789	1032	789	801	837	1002	1083	786	786	819	819	828	1353	822	819	819	741	834	777
120	gp16 terminase DNA packaging enzyme, small subunit	432	435	489	411	432	432	420	432	429	432	426	405	414	438	432	432	495	465	519	534	450	465	495	537	498	495
124	gp17 terminase DNA packaging enzyme large subunit	1644	1653	1683	1683	1644	1653	1647	1650	1650	1650	1653	1089	1650	1659	1650	1650	1833	1842	1902	1803	1839	1842	1833	1830	1824	1836
125	gp18 tail sheath monomer	2190	2250	2010	2034	2253	2250	2232	2262	2259	2253	1908	1413	2247	2421	2253	2250	1980	1992	2040	2016	1992	1992	1980	1995	2001	1983
126	gp19 tail tube monomer	588	591	615	615	585	588	615	609	558	579	618	708	612	684	585	588	492	489	489	501	492	489	492	486	495	492
127	gp20 portal vertex protein of head	1677	1614	1671	1683	1602	1614	1695	1602	1596	1335	1680	1650	1668	1692	1605	1602	1575	1551	1566	1536	1551	1551	1575	1575	1566	1572
129	gp21 prohead core scaffold and protease	651	645	648	648	645	645	645	645	723	645	645	645	651	735	645	645	639|573	633	630	642	633	633	654	648	696	642
408+130	gp22 scaffoldprohead core protein	1101	1002	1053	1053	1047	1050	1179	1038	1026	1041	1185	1131	1098	1089	1038	1020	810	825	795	843	831	825	810	789	795	813
131	gp23 precursor of major head subunit	1413	1389	1368	1368	1398	1389	1407	1374	1374	1380	1407	1398	1404	1407	1395	1377	1566	1590	1605	1545	1590	1590	1566	1575	1587	1569
105	gp25 base plate wedge subunit	402	420	390	396	420	420	393	417	420	420	393	438	402	417	420	420	399	384	423	420	384	384	399	393	387	399
11	gp26 baseplate hub subunit	720	711	699	699	717	708	717	723	714	711	717	720	720	717	711	711	627	615	777	849	606	615	627	558	630	627
5	gp32 ssDNA binding protein	819	924	909	876	921	927	888	948	1026	930	873	999	966	921	981	1020	906	888	909	909	885	888	909	981	969	900
326+2112+ 2096+1667	gp33 late promoter transcription factor	261	261	351	351	390	252	249	249	306	309	249	306	258	336	306	306	339	258	231	297	255	258	339	252	270	339
182	gp41 DNA primase-helicase	1383	1179	1374	1368	1380	1371	1413	1389	1380	1458	1404	1380	1377	1386	1380	1410	1428	1410	1458	1401	1350	1407	1428	1437	1413	1443
178	gp43 DNA polymerase	2496	2481	2487	2487	2487	2526	2493	2499	2499	2499	2490	2502	2496	2505	2499	2499	2697	1176|1497	2760	2553	1497|1523	1176|1497	2697	2706	2679	2712
157	gp44 clamp loader subunit	942	879	942	942	942	1008	948	942	942	942	942	942	951	948	945	942	960	960	966	957	966	960	960	999	975	963
153	gp45 sliding clamp DNA polymerase accessory protein	666	669	669	669	660	660	666	669	666	669	666	654	693	726	660	660	687	672	678	666	672	672	687	642	687	687
143+1865 + 2162	gp46 recombination endonuclease subunit	1713	1719	1722	1722	1722	1722	1731	1722	1722	1722	1731	1722	1713	1725	1722	1722	1683	1713	2319	2238	1713	1713	1689	1704	1683	1689
141+1024	gp47 recombination endonuclease subunit	1041	1032	1047	1047	1035	1035	1050	1023	1035	744|432	954	1044	1050	1044	1035	1035	1020	1068	1029	1041	1068	1068	1020	1020	1026	1020
333+7+1152+ 1620+2019	gp48 baseplate tail tube cap	1161	1071	1326	1332	1041	885	999	1014	1116	762	1140	960	1317	2013	1044	1095	1095	1029	1077	1140	1029	1029	1095	1086	1059	1110
6	gp53 base plate wedge component	726	972	618	303	960	924	663	894	927	900	660	969	669	756	927	894	591	540	567	579	540	540	591	555	555	576
140	gp55 Sigma factor for late transcription	480	477	471	471	474	534	495	468	501	537	495	474	537	480	471	471	558	519	522	513	519	519	558	516	534	558
202	gp61 DNA primase subunit	972	834	999	999	993	969	990	987	825	987	990	996	972	813	993	978	1029	1005	1035	1059	1005	1005	1029	1029	1029	1023
167	gp62 clamp loader subunit	333	405	381	390	402	405	387	405	402	405	387	171	369	384	402	405	564	579	582	489	573	579	564	573	579	564
325+1472	DexA exonuclease A	690	684	591	384	678	684	690	678	684	684	678	672	684	687	678	681	684	666	681	693	666	666	684	675	669	678
203	NrdA ribonucleotide reductase A subunit	2328	2307	2301	2301	2319	2304	2331	2304	2310	2313	2325	2298	2298	2301	2298	2310	2265	1713	1422	2226	1713	1713	2265	2256	2244	2256
311+1531+ 1776	NrdC glutaredoxin	249	177	276	261	237	237|240	246|249	177	240	270|237	246	237|243	327|237	279|249	243	216|237	264	228	273	300	231	228	264	273	282	279
168	RegA translational repressor of early genes	444	420	423	423	426	429	426	450	432	426	426	456	438	462	429	429	369	357	369	381	357	357	369	369	363	378
318+1660 + 2793	Td thymidilate synthase	636	633	660	678	708	810	642	711	645	729	687	711	702	714	702	711	1878	840	834	903	840	840	861	858	1233	861
138	UvsW RNA-DNA + DNA-DNA helicase	1464	1464	663|840	1512	1473	1473	1464	1476	1473	1467	1452	1464	1482	1479	1473	1488	1764	1482	1512	1524	1488	1482	1512	1500	1503	1515
204+1777	NrdB ribonucleotide reductase B subunit	1155	1164	1188	1185	1167	1167	1185	1164	1167	978	1188	1161	1194	1167	1167	1161	1765	972	1131	1125	999	972	1179	1176	1161	1173

*** NOTE: gp51 (25/26), uvsX, uvsY (each in 23/26) and gp59 (22/26) are nearly universal among T4-like phages

Suppl. Table 3: Non-cyano T4-like "core" beyond the T4-core. Numbers listed for each phage are as in Suppl. Table 2.

count	T4-GC #	Function	T4	44RR	Aeh1	KVP40	PHG25	PHG31	RB32	RB43	RB49	RB69	missing in which cyanophages?
1	1642	dCMP deaminase	582	519	549	453	519	519	582	525	507	510	all
2	1475	Dda DNA helicase	1320	1320	1365	1236	1320	1320	1320	1332	1392	1314	all
3	1668	DsbA dsDNA binding protein, late transcription	270	288	288	273	285	288	270	264	276	288	all
4	1586	gp1 dNMP kinase	726	675	690	639	684	675	726	663	657	735	all
5	1593+2021+1074+1204+1328+397+1122+511+50	gp7 baseplate wedge initiator	3099	3060	3492	3498	3057	3060	3099	3084	3087	3099	P-RSM4, P-SSM7, S-PM2, Syn1, Syn9, Syn19, Syn33, S-SSM5, S-SM1
6	398 + 1595	gp9 baseplate wedge tail fiber connector	867	858	927	984	858	858	867	864	855	864	P-SSM2, P-HM1, P-HM2, S-PM2, Syn1, S-SM2
7	1596+2377+2022	gp10 baseplate wedge subunit and tail pin	1809	1815	2157	2247	1815	1815	1806	1818	1803	1815	all
8	1597+2378	gp11 base plate wedge component	660	663	975	705	663	663	660	663	645	660	all
9	1796+1598+2379	gp12 short tail fiber	1584	1401	1311	1422|1539	1398	1401	1551	1392	1401	1551	all
10	1606	gp24 precursor of head vertex subunit	1284	1233	1179	897	1233	1233	1284	1323	1242	1284	all
11	1627	gp30 DNA ligase	1464	1506	1488	1344	1515	1506	1461	1524	1497	1494	all
12	1639	gp31 head assembly cochaperone with GroEL	336	339	411	339	312	339	336	345	324	333	all
13	1669	gp34 long tail fiber, proximal subunit	3870	3669	3711	3771	3666	3669	3870	3660	3741	3834	all
14	1525	gp49 recombination endonuclease VII	474	474	486	456	474	474	474	483	474	474	all
15	1684	gp52 DNA topoisomerase subunit	1329	1668	1317	1287	1635	1665	1329	1164	1365	1326	all
16	1621	gp54 baseplate tail tube initiator	963	858	996	747	858	858	966	864	933	963	all
17	3	gp59 loader of gp41 DNA helicase	654	660	648	651	660	660	654	666	648	528	S-PM2, Syn1, S-SSM7, P-SSM7
18	1464	gp60+39 DNA topoisomerase subunit	1551	1824	1842	1794	1824	1824	1818	1893	1824	1821	all
19	1491	Hypothetical-Protein	165	195	468	477	189	186	165	501	177	279	all
20	1584	Hypothetical with 5' RNA ligase family domain	459	429	495	456	438	429	456	624	465	456	all
21	1630	Hypothetical with DUF1768 domain	459	450	465	465	423	450	459	441	480	414	all
22	1542+2721	NrdC.11 hypothetical protein	1011	975	954	1041	975	975	1014	987	699	990	all
23	1523	NrdD anaerobic NTP reductase large subunit	2851	1827	2115	1836	1827	1827	1818	2124	1863	1818	all
24	1519+2154	NrdH glutaredoxin	309	276	285	240	276	276	309	279	270	273	all
25	1648	PseT polynucleotide 5'-kinase and 3'-phosphatase	906	888	918	918	900	888	909	888	879	900	all
26	1698	RIIA-RIIB membrane-associated	939	1134	1317	1038	942	1134	939	3150	993	936	all
27	870	RNaseH ribonuclease	918	924	921	933	915	924	918	936	948	873	all but S-PM2, Syn1
28	1653	RnlA RNA ligase	1125	1152	1170	1146	1152	1152	1125	1113	1170	1125	all
29	1553	Tk thymidine kinase	582	576	600	585	573	576	582	582	597	582	all
30	1557	Tk.4 hypothetical protein	468	489	609	513	486	489	468	399	456	465	all
31	1559	Vs.1 hypothetical with transglycosylase SLT domain	546	552	525	627	534	552	546	636	591	543	all
32	1599 + 2023	Wac fibritin neck whiskers	1464	1764	3108	1680	1761	1764	1458	2289	1770	1443	all

Suppl. Table 4: Cyano T4-like core genes ***. Numbers listed for each phage are as in Suppl. Table 2.

			Prochlorococcus phages						*Synechococcus* phages									
Count	T4-GC #	GENE DESCRIPTION	P-SSM2	P-SSM4	P-HM1	P-HM2	P-RSM4	P-SSM7	S-PM2	Syn9	Syn19	Syn33	Syn1	S-ShM2	S-SM2	S-SSM7	S-SSM5	S-SM1
1	280	PsbA photosystem II D1 protein	1083	1098	1113	1089	1089	1167	1292	1137	1080	1080	1080	1077	909	1185	1077	1080
2	184	MazG pyrophosphatase	417	402	405	405	402	417	408	402	402	402	408	402	474	426	402	402
3	322	PhoH P-starvation inducible protein	753	774	747	762	762	771	753	765	762	762	756	768	756	768	762	762
4	170	Hsp20 small heat shock protein	459	450	471	498	450	441	411	483	507	492	462	501	492	432	507	447
5	267	Hli03 high-light inducible protein	114|108|219|144	111|207|108	108|114|222	108|114|222	201|165|114	210|147|114	120|195	204|135	255|153	108|219	210|120	210	204	213|156	108|219	255|147
6	150	CobS porphyrin biosynthetic protein	1095	1095	1065	1065	1104	1092	1119	1071	1080	1074	1098	1059	1146	1116	1074	1086
7	15	Virion structural protein	1263	1341	1248	1251	1431	1380	1395	1404	1362	1374	1398	1809	1443	1413	1401	1386
8	4	Hyp. with DUF1825 domain	333	333	321	324	333	330	348	336	327	351	321	336	345	294	333	369
9	146	Hyp. with carboxypeptidase domain	2214	2235	2226	2226	2223	2151	2223	2124	2160	2112	2136	2169	2124	2256	2184	2109
10	190	Hyp. with CTP transerase domain	1275	1152	570	552	1197	1197	1224	1149	1197	1158	570	1179	1227	1383	1197	1200
11	139	Hyp. with Methylamine utilization domain	414	441	429	429	426	444	399	438	438	420	396	447	414	408	432	432
12	101+155	Hyp. with Phytanoyl-CoA-dioxygenase domain	588|642|591|585|567|612|585|498	540	627|588	612	612|558	582	597	585|624	564|633	549	600|387|513	660|648	552	645|648|597	615|555	660|555|540
13	312+443+1092+1149	Hypothetical protein	411	240	237	231	237	246	237	237	234	237	237	243	225	234	234	234
14	176	Hypothetical protein	342	381	426|393	393|423	384	456	468	336	375	351	507	378	396	336	474	381
15	201	Hypothetical protein	294	402	417	417	393	393	495	429	429	381	435	414	402	426	393	429
16	313	Hypothetical protein	624	465	456	456	468	465	435	462	471	468	435	465	483	576	399	471
17	321	Hypothetical protein	327|282	177	243	243	231	231	240	231	243	240	249	231	246	270	231	231
18	49	Hypothetical protein	300	294	300	300	318	300	345	315	315	330	315	315	318	300	312	315
19	71	Hypothetical protein	813	927	1206	1221	996	930	579	963	999	897	1206	921	909	972	891	975
20	112+1330	Hypothetical protein	204	222	222	219	228	258	204	237	246	177	195	237	222	210	228	243
21	142	Hypothetical protein	279	273	285	261	315	249	276	252	261	249	273	252	270	288	261	255
22	152	Hypothetical protein	663	279	411	411	294	312	287	285	279	282	288	291	420	924	279	276
23	250	Hypothetical protein	240	231	291	294	303	249	240	240	243	261	342	243	237	294	297	294
24	198	Hypothetical protein	372	363	372	372	363	363	357	363	363	363	357	363	393	372	363	363
25	43	Hypothetical protein	303	192	276	276	279	171	204	186	174	186	183	174	198	195	180	180

*** NOTE: 9 genes, including 6 hypotheticals (including one with a DUF680 domain), an endonuclease, CP12 and *talC,* are nearly universal cyanophage core genes (missing only in in S-PM2)

Suppl. Table 5: Proteins that are unique to either P-HM1 or P-HM2 phage genome in pairwise comparison of these two co-isolated phages

T4-GC#	***Functional description***	***Genome location***
Unique to P-HM1		
T4-GC171	PurM	156328-156987
T4-GC404	PTOX	173106-173609
T4-GC452	peptidase M15B and M15C	9557-11437
T4-GC277	Hypothetical protein	168032-168250
T4-GC331	Hypothetical protein	2610-3095
T4-GC448	Hypothetical protein	2293-2403
T4-GC467	Hypothetical protein	30252-30749
T4-GC495	Hypothetical protein	39852-40460
T4-GC515	Hypothetical protein	90207-91184
T4-GC516	Hypothetical protein	91187-91579
T4-GC524	Hypothetical protein	112705-112881
T4-GC526	Hypothetical protein	116718-116849
T4-GC527	Hypothetical protein	120914-121168
T4-GC528	Hypothetical protein	127117-127761
T4-GC533	Hypothetical protein	139047-140504
T4-GC543	Hypothetical protein	157906-157799
T4-GC545	Hypothetical protein	158486-159028
T4-GC550	Hypothetical protein	161740-161636
T4-GC552	Hypothetical protein	162104-161988
T4-GC553	Hypothetical protein	162225-162097
T4-GC554	Hypothetical protein	166737-166838
T4-GC558	Hypothetical protein	169174-169284
T4-GC560	Hypothetical protein	173967-174227
T4-GC561	Hypothetical protein	174224-174328
T4-GC563	Hypothetical protein	176365-176544
T4-GC566	Hypothetical protein	178336-178575
T4-GC454	Hypothetical protein	16042-17223
T4-GC461	Hypothetical protein	26834-27736
T4-GC468	Hypothetical protein	31675-31842
T4-GC469	Hypothetical protein	31832-32032
T4-GC470	Hypothetical protein	32164-32012
T4-GC473	Hypothetical protein	33188-33328
T4-GC480	Hypothetical protein	35023-35205
T4-GC483	Hypothetical protein	37113-36985
T4-GC486	Hypothetical protein	38194-38307
T4-GC488	Hypothetical protein	38631-38732
T4-GC491	Hypothetical protein	39235-39375
T4-GC493	Hypothetical protein	39559-39756
T4-GC494	Hypothetical protein	39749-39892
T4-GC499	Hypothetical protein	41175-41390
T4-GC502	Hypothetical protein	42316-42492
T4-GC213	Hypothetical protein	157848-158333

(continued on next page)

Unique to P-HM2

T4-GC588	putative restriction endonuclease	133838-134473
T4-GC568	endodeoxyribonuclease	2798-3409
T4-GC573	peptidase M15B and M15C	10357-12189
T4-GC575	Kelch repeat-containing protein	27671-28591
T4-GC432	Hypothetical protein	116970-117137
T4-GC587	Hypothetical protein	124954-125154
T4-GC419	Hypothetical protein	143644-145134
T4-GC590	Hypothetical protein	161186-161338
T4-GC591	Hypothetical protein	161341-161601
T4-GC592	Hypothetical protein	168030-168140
T4-GC593	Hypothetical protein	168739-168846
T4-GC594	Hypothetical protein	169141-169257
T4-GC595	Hypothetical protein	171570-171722
T4-GC596	Hypothetical protein	176033-176221
T4-GC597	Hypothetical protein	176384-176512
T4-GC598	Hypothetical protein	176496-176606
T4-GC599	Hypothetical protein	178664-178870
T4-GC355	Hypothetical protein	178936-179310
T4-GC600	Hypothetical protein	182992-183117
T4-GC574	Hypothetical protein	19368-19865
T4-GC567	Hypothetical protein	2410-2718
T4-GC576	Hypothetical protein	33484-33230
T4-GC577	Hypothetical protein	34793-35071
T4-GC578	Hypothetical protein	35501-35602
T4-GC579	Hypothetical protein	36219-36338
T4-GC580	Hypothetical protein	39242-39370
T4-GC581	Hypothetical protein	39679-39795
T4-GC582	Hypothetical protein	40493-40708
T4-GC583	Hypothetical protein	41516-41644
T4-GC584	Hypothetical protein	42554-42727
T4-GC569	Hypothetical protein	4849-4953
T4-GC570	Hypothetical protein	4919-5023
T4-GC571	Hypothetical protein	6106-5957
T4-GC572	Hypothetical protein	6158-6289
T4-GC351	Hypothetical protein	90401-91570
T4-GC352	Hypothetical protein	91573-92319
T4-GC353	Hypothetical protein	92333-93100
T4-GC354	Hypothetical protein	93054-94793
T4-GC585	Hypothetical protein	94888-95793
T4-GC586	Hypothetical protein	95790-96236

Suppl. Table 6: *Synechococcus* phage enriched proteins. Numbers listed for each phage are as in Suppl. Table 2.

		Prochlorococcus						*Synechecoccus*									
T4-GC #	GENE DESCRIPTION	P-SSM2	P-SSM4	P-HM1	P-HM2	P-RSM4	P-SSM7	S-PM2	Syn9	Syn19	Syn33	Syn1	S-ShM2	S-SM2	S-SSM7	S-SSM5	S-SM1
881	Hypothetical protein	--	--	--	--	--	--	--	207	--	204	150	--	--	--	--	213
937	Hypothetical protein	--	--	--	--	--	--	--	165	--	168	177	168	279	--	--	--
957	Hypothetical protein	--	--	--	--	--	--	--	--	198	231	--	--	--	--	198	168
810	Hypothetical protein	--	--	--	--	--	--	204	219	--	--	--	--	231	--	--	--
927	Hypothetical protein	--	--	--	--	--	--	--	201	--	222	--	222	--	--	--	--
924	Hypothetical protein	--	--	--	--	--	--	--	165	165	168	--	--	--	--	--	162
931	Hypothetical protein	--	--	--	--	--	--	--	207	--	207	198	225	--	--	--	--
1011	Hypothetical protein	--	--	--	--	--	--	--	--	--	204	--	201	210	--	--	--
838	Hypothetical protein	--	--	--	--	--	--	219	273	339	237	279	270	297	--	--	357
1013	Hypothetical protein	--	--	--	--	--	--	--	--	--	279	243	423	--	--	--	--
730	Hypothetical protein	--	--	--	--	--	--	273	--	--	--	276	--	270	--	--	--
744	Hypothetical protein	--	--	--	--	--	--	288	237	219	297	--	270	342	--	--	213
751	Hypothetical protein	--	--	--	--	--	--	306	--	--	357	231	--	--	--	--	--
942	Hypothetical protein	--	--	--	--	--	--	--	588	591	573	690	483	573	--	--	--
945	Hypothetical protein	--	--	--	--	--	--	--	654	669	678	672	690	753	672	678	678
920	6PGDH = gnd	--	--	--	--	--	--	--	1038	1038	--	1038	1023	1038	--	1041	1041
921+1021	G6PDH = zwf	--	--	--	--	--	--	--	1446	1440	276	303	306	1440	--	1443	1437
1035	SAICAR synthetase - purine synthesis	--	--	--	--	--	--	--	--	--	699	699	699	--	--	--	--
969	virion structural protein	--	--	--	--	--	--	--	--	19452	--	--	--	--	--	18516	18543
1038	Hyp. w/ PA14 carbohydrate binding domain	--	--	--	--	--	--	--	--	--	2214	--	--	4275	--	2127	2109
928	Hyp. W/ DUF1583 domain	--	--	--	--	--	--	--	231	--	234	--	255	--	--	--	237
876	Hypothetical protein	--	--	--	--	--	--	--	183	237	201	201	210	183	195	--	219
884	Hypothetical protein	--	--	--	--	--	--	--	222	162	159	150	--	165	--	159	165
922	Hypothetical protein	--	--	--	--	--	--	--	180	183	153	162	--	159	--	--	204
987	Hypothetical protein	--	--	--	--	--	--	--	--	219	--	222	228	189	--	219	234
988	Hypothetical protein	--	--	--	--	--	--	--	--	249	--	--	273	225	--	240	240
900	Hypothetical protein	--	--	--	--	--	--	--	189	240	252	--	231	--	--	--	--
903	Hypothetical protein	--	--	--	--	--	--	--	168	--	159	--	--	--	--	168	189
919	Hypothetical protein	--	--	--	--	--	--	--	129	219	--	--	--	120	--	--	171
923	Hypothetical protein	--	--	--	--	--	--	--	135	--	--	--	252	225	--	--	135
934	Hypothetical protein	--	--	--	--	--	--	--	585	585	621	--	627	--	--	--	--
943	Hypothetical protein	--	--	--	--	--	--	--	351	--	--	--	--	399	--	399	477
948	Hypothetical protein	--	--	--	--	--	--	--	180	--	180	177	219	--	--	--	--
1010	Hypothetical protein	--	--	--	--	--	--	--	--	--	135	144	144	--	--	--	--
1021	Hypothetical protein	--	--	--	--	--	--	--	--	--	276	303	306	--	--	--	--
1026	Hypothetical protein	--	--	--	--	--	--	--	--	--	300	396	429	--	--	--	--
1036	Hypothetical protein	--	--	--	--	--	--	--	--	--	273	273	270	--	--	--	--
755	Hypothetical protein	--	--	--	--	--	--	396	--	--	--	--		--	--	357	363
835	Hypothetical protein	--	--	--	--	--	--	543	546	--	--	--	546	--	--	--	
901	Hypothetical protein	--	--	--	--	--	--	--	195	198	--	--		--	--	--	216
918	Hypothetical protein	--	--	--	--	--	--	--	240	--	216	--	225	--	--	--	--
930	Hypothetical protein	--	--	--	--	--	--	--	120	--	309	--	189	--	--	--	--
953	Hypothetical protein	--	--	--	--	--	--	--	--	204	--	--	--	--	--	216	180
955	Hypothetical protein	--	--	--	--	--	--	--	--	264	--	--	--	--	--	285	264
956	Hypothetical protein	--	--	--	--	--	--	--	--	159	--	--	--	--	--	141	141
958	Hypothetical protein	--	--	--	--	--	--	--	--	177	--	--	--	--	--	177	177
959	Hypothetical protein	--	--	--	--	--	--	--	--	339	--	--	--	--	--	231	222
964	Hypothetical protein	--	--	--	--	--	--	--	--	138	--	--	--	--	--	138	264

Suppl. Table 7: *Prochlorococcus* phage enriched proteins. Numbers listed for each phage are as in Suppl. Table 2.

T4-GC #	GENE DESCRIPTION	P-SSM2	P-SSM4	P-HM1	P-HM2	P-RSM4	P-SSM7	S-PM2	Syn9	Syn19	Syn33	Syn1	S-ShM2	S-SM2	S-SSM7	S-SSM5	S-SM1
163	***Possible PsbN photosystem protein******	303	225	210	210	150	306	--	--	--	--	--	--	--	--	--	--
436	Hli04_PSSM4	--	135	135	201	219	--	--	--	--	--	--	--	--	--	--	--
413	PcyA, phycocyanobilin biosynthesis protein	--	690	--	--	729	717	--	--	--	--	--	--	--	--	--	--
285	***Hypothetical protein******	282	288	291	288	288	264	--	--	--	--	--	--	--	--	--	--
429	Hypothetical protein	--	210	204	204	207	--	--	--	--	--	--	--	--	--	--	--
437	Hypothetical protein	--	174	180	177	--	201	--	--	--	--	--	--	--	--	--	--
485	Hypothetical protein	--	--	177	177	183	204	--	--	--	--	--	--	--	--	--	--
95	Hypothetical protein	252	339	--	--	252	201	--	--	--	--	--	--	--	--	--	--
367	Hypothetical protein	--	159	207	186	--	--	--	--	--	--	--	--	--	--	--	--
387	Hypothetical protein	--	234	249	249	--	--	--	--	--	--	--	--	--	--	--	--
391	Hypothetical protein	--	174	177	177	--	--	--	--	--	--	--	--	--	--	--	--
423	Hypothetical protein	--	216	219	222	--	--	--	--	--	--	--	--	--	--	--	--
466	Hypothetical protein	--	--	264	264	348	--	--	--	--	--	--	--	--	--	--	--
496	Hypothetical protein	--	--	198	195	171	--	--	--	--	--	--	--	--	--	--	--
506	Hypothetical protein	--	--	261	210	201	--	--	--	--	--	--	--	--	--	--	--
542	Hypothetical protein	--	--	126	126	--	189	--	--	--	--	--	--	--	--	--	--
596	Hypothetical protein	--	--	--	189	228	210	--	--	--	--	--	--	--	--	--	--
79	Hypothetical protein	216	168	--	--	162	--	--	--	--	--	--	--	--	--	--	--
82	***Hypothetical protein***	288	225	264	249	249	234	--	--	--	--	--	--	--	297	243	--
224	***Hypothetical protein***	225	159	183	162	213	207	--	--	--	--	--	--	--	207	--	--

Suppl. Table 8: Summary of cyano T4 proteomics experiments. Comparative proteomics = experimentally determined protein content in purified virus particles to determine the structural proteins in three sequenced T4-like virus genomes. An “Y” means the protein was detected, “--“ means the protein is annotated in the genome but no peptides were detected, “NP” means the protein is not present in the genome, “counts” are the number of peptide fragments detected per protein, "copy # in T4" refers to the biochemically and ultrastructurall determined copy number ofproteins in the coliphage T4 particle. Ten of these proteins, in italics, have similar distributions among 9 cyanophages and may be functionally linked.

		Proteomic data						Genomic distribution of the genes, sizes in nucleotides														
		S-SM1[1]		S-PM2[2]		SYN9[3]																
T4-GC #	DEFINITION	DETECTED	COUNTS	DETECTED	COUNTS	DETECTED	COPY # in T4	PSSM2	PSSM4	P-HM1	P-HM2	P-RSM4	P-SSM7	S-PM2	SYN9	Syn19	Syn33	Syn1	S-ShM2	S-SM2	S-SSM7	S-SSM5	S-SM1
125	gp18	Y	70	Y	18	Y	138	2190	2250	2010	2034	2253	2250	2232	2262	2259	2253	1908	1413	2247	2421	2253	2250
126	gp19	Y	73	Y	3	Y	144	588	591	615	615	585	588	615	609	558	579	618	708	612	684	585	588
127	gp20	Y	18	Y	5	Y	12	1677	1614	1671	1683	1602	1614	1695	1602	1596	1335	1680	1650	1668	1692	1605	1602
131	gp23	Y	119	Y	22	Y	960	1413	1389	1368	1368	1398	1389	1407	1374	1374	1380	1407	1398	1404	1407	1395	1377
106	gp6	Y	25	Y	3	Y	12	1944	1989	1851	1851	2028	2031	1809	2031	2025	2031	1809	2133	2112	1875	2028	2019
116	gp13	Y	13	Y	5	-	10	846	807	816	816	810	807	831	807	807	807	831	807	846	819	807	807
118	gp15	Y	15	Y	5	-	6	843	1032	1005	1005	786	789	801	789	1032	786	801	837	1002	1083	786	786
133	gp3	Y	18	Y	2	-	6	534	552	558	546	549	549	510	573	570	549	506	567	576	561	549	549
408	gp22	Y	22	Y	3	-	115	1101	1002	1053	1053	1047	1050	1179	1038	1026	1041	1185	1131	1098	1089	1038	1020
333	gp48	Y	13	Y	2	-	6	1161	1071	1326	1332	1041	885	999	1014	1116	762	1140	960	1317	2013	1044	1095
108	gp8	Y	21	Y	19	Y	12	1602	1533	1515	1515	1539	1533	1905	1533	1533	1533	1884	1533	1551	1596	1533	1533
402	*structural protein*	*Y*	*14*	*NP*	-	*Y*	*NP*	--	*7974*	--	--	*6711*	*6543*	--	*5340*	*6489*	*5724*	--	*6687*	--	--	*5949*	*5967*
346	*fiber*	*Y*	*16*	*NP*	-	*Y*	*NP*	--	*3957*	--	--	*3963*	*3966*	--	*3954*	*3978*	*3285*	--	*3816*	--	--	*3963*	*3963*
344	*Structural protein*	*Y*	*16*	*NP*	-	*Y*	*NP*	--	*1323*	--	--	*1326*	*1323*	--	*1329*	*1323*	*1326*	--	*1320*	--	--	*1326*	*1326*
398	*gp9*	*Y*	*25*	*NP*	-	*NP*	*18*	--	*1230*	--	--	*1236*	*1227*	--	*1221*	*1239*	*1221*	--	*945*	--	*735*	*1227*	*1227*
403	*Structural protein*	*Y*	*60*	*NP*	-	*NP*	*NP*	--	*534*	--	--	*537*	*540*	--	--	*531*	*537*	--	--	--	--	*534*	*537*
334	*Structural protein*	*Y*	*34*	*NP*	-	-	*NP*	--	*990*	--	--	*795*	*837*	--	*846*	*810*	*846*	--	*606*	--	--	--	*795*
347	*Structural protein*	*Y*	*32*	*NP*	-	-	*NP*	--	*417*	--	--	*417*	*441*	--	*918*	*417*	*417*	--	*435*	--	--	*417*	*423*
399	*Structural protein*	*Y*	*8*	*NP*	-	-	*NP*	--	*4332*	--	--	*4143*	*4128*	--	*4161*	*4200*	--	--	*3555*	--	--	*4161*	*4164*
426	*Structural protein*	*Y*	*17*	*NP*	-	-	*NP*	--	*540*	--	--	--	*582*	--	*525*	*558*	*558*	--	*861*	--	--	*513*	*594*
345	*Structural protein*	-	-	-	-	*Y*	*NP*	--	*1485*	--	--	*1452*	*1455*	--	*1506*	*1491*	*1488*	--	*1479*	--	--	*1449*	*1485*
512	Structural protein	NP	-	Y	2	NP	NP	--	--	534	534	--	--	507	--	--	--	513	--	--	--	--	--
537	Structural protein	NP	-	Y	5	NP	NP	--	--	864	864	--	--	951	--	--	--	858	--	--	--	--	--
425	S-layer domain	Y	17	NP	-	NP	NP	--	573	--	--	573	633	--	--	--	--	--	--	--	564	573	573
400	Structural protein	Y	17	NP	-	Y	NP	--	7248	6654	6669	7467	7605	--	7176	7293	7209	--	7257	7560	--	7458	7173
15	Structural protein	Y	12	-	-	Y	NP	1263	1410	1248	1251	1431	1380	1395	1404	1362	1374	1398	1809	1443	1413	1401	1386
6	gp53	Y	6	-	-	-	6	726	972	618	303	960	924	663	894	927	900	660	969	669	756	927	894
190	Cytidyldyltransfera	Y	25	-	-	-	NP	1275	1152	570	552	1197	1197	1224	1149	1197	1158	570	1179	1227	1383	1197	1200
117	gp14	Y	5	-	-	-	5	1413	927	1329	1329	933	927	879	1173	921	1173	879	1164	1455	2289	927	927
129	gp21	Y	8	-	-	-	3	651	645	648	648	645	645	645	645	723	645	645	645	651	735	645	645
105	gp25	Y	15	-	-	-	6	402	420	390	396	420	420	393	417	420	420	393	438	402	417	420	420
9	gp4	Y	4	-	-	-	??	438	426	435	462	426	420	438	420	420	420	342	438	444	480	426	474
340	gp5	Y	11	-	-	-	3	2259	2310	870	873	2541	2553	2946	2508	2385	2580	2922	2484	1848	3027	2544	2553
119	Structural protein	Y	13	NP	-	-	NP	816	1386	702	894	1098	519	--	1332	1164	1371	276	480	2127	2985	1206	1488
607	Lysozyme murein	Y	16	NP	-	NP	NP	--	--	--	--	2802	--	--	--	--	--	--	--	--	--	2862	2757
1038	PA14 domain	Y	5	NP	-	NP	NP	--	--	--	--	--	--	--	--	--	2214	--	--	4275	--	2127	2109
1426	Structural protein	Y	17	NP	-	NP	NP	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	2757
1428	Structural protein	Y	25	NP	-	NP	NP	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	948
1453	Structural protein	Y	12	NP	-	NP	NP	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	834
1454	Structural protein	Y	10	NP	-	NP	NP	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	1422
1455	Structural protein	Y	7	NP	-	NP	NP	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	489
641	Structural protein	Y	9	NP	-	NP	NP	--	--	--	--	2640	--	--	--	--	2553	--	--	--	--	2589	2592
642	Structural protein	Y	5	NP	-	-	NP	--	--	--	--	609	582	--	522	--	--	--	522	--	--	582	552
969	Structural protein	Y	8	NP	-	NP	NP	--	--	--	--	--	--	--	--	19452	--	--	--	--	--	18516	18543
735	Structural protein	NP	-	Y	2	NP	NP	--	--	--	--	--	--	1350	--	--	--	1704	--	--	--	--	--
737	Structural protein	NP	-	Y	3	NP	NP	--	--	--	--	--	--	528	--	--	--	531	--	--	--	--	--
827	Structural protein	NP	-	Y	3	NP	NP	--	--	--	--	--	--	1704	--	--	--	1851	--	--	--	--	--
739	Structural protein	NP	-	Y	3	NP	NP	--	--	--	--	--	--	1749	--	--	--	--	--	--	--	--	--
829	Structural protein	NP	-	Y	9	NP	NP	--	--	--	--	--	--	3756	--	--	--	--	--	--	--	--	--
831	Structural protein	NP	-	Y	5	NP	NP	--	--	--	--	--	--	921	--	--	--	--	--	--	--	--	--
832	Structural protein	NP	-	Y	2	NP	NP	--	--	--	--	--	--	984	--	--	--	--	--	--	--	--	--
833	Structural protein	NP	-	Y	2	NP	NP	--	--	--	--	--	--	1140	--	--	--	--	--	--	--	--	--
864	Structural protein	NP	-	Y	5	NP	NP	--	--	--	--	--	--	3177	--	--	--	--	--	--	--	--	--
267	Putative HLIP	-	-	Y	2	-	NP	114| 108| 219| 144	111| 207| 108	108| 114| 222	108| 114| 222	201| 165| 114	210| 147| 114	120| 195	204| 135	255| 153	108| 219	210| 120	210	204	213| 156	108| 219	255| 147

References

Alemayehu, D., Ross, R. P., O'Sullivan, O., Coffey, A., Stanton, C., Fitzgerald, G. F., & McAuliffe, O. (2009). Genome of a virulent bacteriophage Lb338-1 that lyses the probiotic Lactobacillus paracasei cheese strain. *Gene*, 448(1):29–39.

Alm, E. J., Huang, K. H., Price, M. N., Koche, R. P., Keller, K., Dubchak, I. L., & Arkin, A. P. (2005). The MicrobesOnline Web site for comparative genomics. *Genome Res*, 15(7):1015–22.

Amla, D., Rowell, P., & Stewart, W. (1987). Metabolic changes associated with cyanophage N-1 infection of the cyanobacterium Nostoc muscorum. *Arch Microbiol*, 148:321–7.

Anisimova, M. & Gascuel, O. (2006). Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Syst Biol*, 55(4):539–52.

Arias, M., Lenardon, S., & Taleisnik, E. (2003). Carbon metabolism alterations in sunflower plants infected with the Sunflower chlorotic mottle virus. *J Phytopathol*, 151(5):267–273.

Arnold, K., Bordoli, L., Kopp, J., & Schwede, T. (2006). The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*, 22(2):195–201.

Bagby, S. C. (2009). *Life in a drop of water*. PhD Thesis, Department of Biology, Massachusetts Institute of Technology.

Bailey, S., Melis, A., Mackey, K. R. M., Cardol, P., Finazzi, G., van Dijken, G., Berg, G. M., Arrigo, K., Shrager, J., & Grossman, A. (2008). Alternative photosynthetic electron flow to oxygen in marine Synechococcus. *Biochim Biophys Acta*, 1777(3):269–76.

Ballou, C. E. (1963). Preparation and droperties of D-erythrose 4-phosphate. *Methods Enzymol*, 6:479–484.

Balogh, A., Borbély, G., Cséke, C., Udvardy, J., & Farkas, G. L. (1979). Virus infection affects the molecular properties and activity of glucose-6-P dehydrogenase in Anacystis nidulans, a Cyanobacterium. Novel aspect of metabolic control in a phage-infected cell. *FEBS Lett*, 105(1):158–62.

Banki, K., Hutter, E., Colombo, E., Gonchoroff, N. J., & Perl, A. (1996). Glutathione levels and sensitivity to apoptosis are regulated by changes in transaldolase expression. *J Biol Chem*, 271(51):32994–3001.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. (2008). GenBank. *Nucleic Acids Res*, 36(Database issue):D25–30.

Berg, J. M., Tymoczko, J. L., & Stryer, L. (2007). *Biochemistry*. Freeman.

Bergh, O., Børsheim, K. Y., Bratbak, G., & Heldal, M. (1989). High abundance of viruses found in aquatic environments. *Nature*, 340(6233):467–8.

Bergmeyer, H. U., Gawehn, K., & Grassl, M. (1974). Enzymes as biochemical reagents. In: *Methods of Enzymatic Analysis*, H. U. Bergmeyer, ed., volume 1, pages 425–522. Academic Press.

Beumer, A. & Robinson, J. B. (2005). A broad-host-range, generalized transducing phage (SN-T) acquires 16S rRNA genes from different genera of bacteria. *Appl Environ Microbiol*, 71(12):8301–4.

Birge, E. (2006). *Bacterial and Bacteriophage Genetics*. Springer.

Blankenship, R. E. (2002). *Molecular Mechanisms of Photosynthesis*. Blackwell.

Bouman, H. A., Ulloa, O., Scanlan, D. J., Zwirglmaier, K., Li, W. K. W., Platt, T., Stuart, V., Barlow, R., Leth, O., Clementson, L., Lutz, V., Fukasawa, M., Watanabe, S., & Sathyendranath, S. (2006). Oceanographic basis of the global surface distribution of Prochlorococcus ecotypes. *Science*, 312(5775):918–21.

Bradford, M. M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem*, 72:248–54.

Breitbart, M., Thompson, L. R., Suttle, C. A., & Sullivan, M. B. (2007). Exploring the vast diversity of marine viruses. *Oceanography*, 20(2):135–139.

Brown, C., Lawrence, J., & Campbell, D. (2006). Are phytoplankton population density maxima predictable through analysis of host and viral genomic DNA content? *J Mar Biol Assoc UK*, 86(03):491–498.

Brussaard, C. P. D., Wilhelm, S. W., Thingstad, F., Weinbauer, M. G., Bratbak, G., Heldal, M., Kimmance, S. A., Middelboe, M., Nagasaki, K., Paul, J. H., Schroeder, D. C., Suttle, C. A., Vaqué, D., & Wommack, K. E. (2008). Global-scale processes with a nanoscale drive: the role of marine viruses. *ISME J*, 2(6):575–8.

Chen, F. & Lu, J. (2002). Genomic sequence and evolution of marine cyanophage P60: a new insight on lytic and lysogenic phages. *Appl Environ Microbiol*, 68(5):2589–94.

Chen, X., Mathews, C. K., Wheeler, L. J., Maley, G., Maley, F., & Coombs, D. H. (1995). An immunoblot assay reveals that bacteriophage T4 thymidylate synthase and dihydrofolate reductase are not virion proteins. *J Virol*, 69(4):2119–25.

Choi, K. H., Lai, V., Foster, C. E., Morris, A. J., Tolan, D. R., & Allen, K. N. (2006). New superfamily members identified for Schiff-base enzymes based on verification of catalytically essential residues. *Biochemistry*, 45(28):8546–55.

Clokie, M. R. J. & Mann, N. H. (2006). Marine cyanophages and light. *Environ Microbiol*, 8(12):2074–82.

Clokie, M. R. J., Shan, J., Bailey, S., Jia, Y., Krisch, H. M., West, S., & Mann, N. H. (2006). Transcription of a 'photosynthetic' T4-type phage during infection of a marine cyanobacterium. *Environ Microbiol*, 8(5):827–35.

Coleman, M. L., Sullivan, M. B., Martiny, A. C., Steglich, C., Barry, K., Delong, E. F., & Chisholm, S. W. (2006). Genomic islands and the ecology and evolution of Prochlorococcus. *Science*, 311(5768):1768–70.

Collaborative Computational Project, N. . (1994). The CCP4 suite: programs for protein crystallography. *Acta Crystallogr D Biol Crystallogr*, 50(Pt 5):760–3.

Comeau, A. M., Bertrand, C., Letarov, A., Tétart, F., & Krisch, H. M. (2007). Modular architecture of the T4 phage superfamily: a conserved core genome and a plastic periphery. *Virology*, 362(2):384–96.

Cooley, J. W., Howitt, C. A., & Vermaas, W. F. (2000). Succinate:quinol oxidoreductases in the cyanobacterium synechocystis sp. strain PCC 6803: presence and function in metabolism and electron transport. *J Bacteriol*, 182(3):714–22.

Cremona, T., Kowal, J., & Horecker, B. L. (1965). The mechanism of action of aldolases. XI. Activation by aromatic sulfhydryl reagents and beta-elimination of selected thiol groups. *Proc Natl Acad Sci USA*, 53(6):1395–403.

Cséke, C., Balogh, A., & Farkas, G. L. (1981). Redox modulation of glucose-6-P dehydrogenase in Anacystis nidulans and its 'uncoupling' by phage infection. *FEBS Lett*, 126(1):85–8.

Cséke, C. S. & Farkas, G. L. (1979). Effect of light on the attachment of cyanophage AS-1 to Anacystis nidulans. *J Bacteriol*, 137(1):667–9.

Dammeyer, T., Bagby, S. C., Sullivan, M. B., Chisholm, S. W., & Frankenberg-Dinkel, N. (2008). Efficient phage-mediated pigment biosynthesis in oceanic cyanobacteria. *Curr Biol*, 18(6):442–8.

DeLong, E. F., Preston, C. M., Mincer, T., Rich, V., Hallam, S. J., Frigaard, N., Martinez, A., Sullivan, M. B., Edwards, R., Brito, B. R., Chisholm, S. W., & Karl, D. M. (2006). Community genomics among stratified microbial assemblages in the ocean's interior. *Science*, 311(5760):496–503.

Depew, R. E. & Cozzarelli, N. R. (1974). Genetics and physiology of bacteriophage T4 3'-phosphatase: evidence for involvement of the enzyme in T4 DNA metabolism. *J Virol*, 13(4):888–97.

DuRand, M., Olson, R., & Chisholm, S. (2001). Phytoplankton population dynamics at the Bermuda Atlantic Time-series station in the Sargasso Sea. *Deep-Sea Research II*, 48(8-9):1983–2003.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32(5):1792–7.

Emsley, P. & Cowtan, K. (2004). Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr*, D60:2126–32.

Evans, C., Malin, G., Mills, G., & Wilson, W. (2006). Viral infection of Emiliania huxleyi (Prymnesiophyceae) leads to elevated production of reactive oxygen species. *J Phycol*, 42:1040–1047.

Fahnert, B., Lilie, H., & Neubauer, P. (2004). Inclusion bodies: formation and utilisation. *Adv Biochem Eng Biotechnol*, 89:93–142.

Fareed, G. C. & Richardson, C. C. (1967). Enzymatic breakage and joining of deoxyribonucleic acid. II. The structural gene for polynucleotide ligase in bacteriophage T4. *Proc Natl Acad Sci USA*, 58(2):665–72.

Fernández-González, B., Martínez-Férez, I. M., & Vioque, A. (1998). Characterization of two carotenoid gene promoters in the cyanobacterium Synechocystis sp. PCC 6803. *Biochim Biophys Acta*, 1443(3):343–51.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., & Merrick, J. M. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 269(5223):496–512.

Garcia-Pichel, F., Belnap, J., Neuer, S., & Schanz, F. (2003). Estimages of global cyanobacterial biomass and its distribution. *Algol. Stud.*, 109:213–27.

Gleason, F. K. (1996). Glucose-6-phosphate dehydrogenase from the cyanobacterium, Anabaena sp. PCC 7120: purification and kinetics of redox modulation. *Arch Biochem Biophys*, 334(2):277–83.

Gold, L. M. & Schweiger, M. (1969). The initiation of T4 deoxyribonucleic acid-dependent beta-glucosyltransferase synthesis in vitro. *J Biol Chem*, 244(19):5100–4.

Guindon, S. & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, 52(5):696–704.

Hadas, H., Einav, M., Fishov, I., & Zaritsky, A. (1997). Bacteriophage T4 development depends on the physiology of its host Escherichia coli. *Microbiology*, 143 (Pt 1):179–85.

Hagen, K. D. & Meeks, J. C. (2001). The unique cyanobacterial protein OpcA is an allosteric effector of glucose-6-phosphate dehydrogenase in Nostoc punctiforme ATCC 29133. *J Biol Chem*, 276(15):11477–86.

He, Q., Dolganov, N., Bjorkman, O., & Grossman, A. R. (2001). The high light-inducible polypeptides in Synechocystis PCC6803. Expression and function in high light. *J Biol Chem*, 276(1):306–14.

Heinrich, P. C., Morris, H. P., & Weber, G. (1976). Behavior of transaldolase (EC 2.2.1.2) and transketolase (EC 2.2.1.1) activities in normal, neoplastic, differentiating, and regenerating liver. *Cancer Res*, 36(9 pt.1):3189–97.

Henn, M. R., Sullivan, M. B., Stange-Thomann, N., Osburne, M. S., Berlin, A. M., Kelly, L., Yandava, C., Kodira, C., Zeng, Q., Weiand, M., Sparrow, T., Saif, S., Giannoukos, G., Young, S. K., Nusbaum, C., Birren, B. W., & Chisholm, S. W. (2010). Analysis of high-throughput sequencing and annotation strategies for phage genomes. *PLoS ONE*, 5(2):e9083.

Horecker, B. L. (2002). The pentose phosphate pathway. *J Biol Chem*, 277(50):47965–71.

Horecker, B. L., Pontremoli, S., Ricci, C., & Cheng, T. (1961). On the nature of the transaldolase-dihydroxyacetone complex. *Proc Natl Acad Sci USA*, 47:1949–55.

Horie, M., Honda, T., Suzuki, Y., Kobayashi, Y., Daito, T., Oshida, T., Ikuta, K., Jern, P., Gojobori, T., Coffin, J. M., & Tomonaga, K. (2010). Endogenous non-retroviral RNA virus elements in mammalian genomes. *Nature*, 463(7277):84–7.

Jia, J., Schörken, U., Lindqvist, Y., Sprenger, G. A., & Schneider, G. (1997). Crystal structure of the reduced Schiff-base intermediate complex of transaldolase B from Escherichia coli: mechanistic implications for class I aldolases. *Protein Sci*, 6(1):119–24.

Johnson, Z. I., Zinser, E. R., Coe, A., McNulty, N. P., Woodward, E. M. S., & Chisholm, S. W. (2006). Niche partitioning among Prochlorococcus ecotypes along ocean-scale environmental gradients. *Science*, 311(5768):1737–40.

Jordan, A. & Reichard, P. (1998). Ribonucleotide reductases. *Annu Rev Biochem*, 67:71–98.

Kao, C. C., Green, S., Stein, B., & Golden, S. S. (2005). Diel infection of a cyanobacterium by a contractile bacteriophage. *Appl Environ Microbiol*, 71(8):4276–9.

Kettler, G. C., Martiny, A. C., Huang, K., Zucker, J., Coleman, M. L., Rodrigue, S., Chen, F., Lapidus, A., Ferriera, S., Johnson, J., Steglich, C., Church, G. M., Richardson, P., & Chisholm, S. W. (2007). Patterns and implications of gene gain and loss in the evolution of Prochlorococcus. *PLoS Genet*, 3(12):e231.

Kim, S. K., Makino, K., Amemura, M., Shinagawa, H., & Nakata, A. (1993). Molecular analysis of the phoH gene, belonging to the phosphate regulon in Escherichia coli. *J Bacteriol*, 175(5):1316–24.

Kimber, M. S., Vallee, F., Houston, S., Necakov, A., Skarina, T., Evdokimova, E., Beasley, S., Christendat, D., Savchenko, A., Arrowsmith, C. H., Vedadi, M., Gerstein, M., & Edwards, A. M. (2003). Data mining crystallization databases: knowledge-based approaches to optimize protein crystal screens. *Proteins*, 51(4):562–8.

Klumpp, J., Dorscht, J., Lurz, R., Bielmann, R., Wieland, M., Zimmer, M., Calendar, R., & Loessner, M. J. (2008). The terminally redundant, nonpermuted genome of Listeria bacteriophage A511: a model for the SPO1-like myoviruses of gram-positive bacteria. *J Bacteriol*, 190(17):5753–65.

Krissinel, E. & Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *J Mol Biol*, 372(3):774–97.

Kyte, J. & Doolittle, R. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, 157(1):105–132.

Laemmli, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*, 227(5259):680–5.

Laskowski, R. A., MacArthur, M. W., Moss, D. S., & Thornton, J. M. (1993). AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J Appl Cryst*, 26(2):283–91.

Latifi, A., Ruiz, M., & Zhang, C.-C. (2009). Oxidative stress in cyanobacteria. *FEMS Microbiol Rev*, 33(2):258–78.

Lebreton, S., Andreescu, S., Graciet, E., & Gontero, B. (2006). Mapping of the interaction site of CP12 with glyceraldehyde-3-phosphate dehydrogenase from Chlamydomonas reinhardtii: functional consequences for glyceraldehyde-3-phosphate dehydrogenase. *FEBS J*, 273(14):3358–69.

Leiman, P. G., Kanamaru, S., Mesyanzhinov, V. V., Arisaka, F., & Rossmann, M. G. (2003). Structure and morphogenesis of bacteriophage T4. *Cell Mol Life Sci*, 60(11):2356–70.

Leonardo, M. R., Dailly, Y., & Clark, D. P. (1996). Role of NAD in regulating the adhE gene of Escherichia coli. *J Bacteriol*, 178(20):6013–8.

Li, W., Rao, D., Harrison, W., Smith, J., Cullen, J., Irwin, B., & Platt, T. (1983). Autotrophic picoplankton in the tropical ocean. *Science*, 219(4582):292–295.

Lindell, D., Jaffe, J. D., Coleman, M. L., Futschik, M. E., Axmann, I. M., Rector, T., Kettler, G., Sullivan, M. B., Steen, R., Hess, W. R., Church, G. M., & Chisholm, S. W. (2007). Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. *Nature*, 449(7158):83–6.

Lindell, D., Jaffe, J. D., Johnson, Z. I., Church, G. M., & Chisholm, S. W. (2005). Photosynthesis genes in marine viruses yield proteins during host infection. *Nature*, 438(7064):86–9.

Lindell, D., Sullivan, M. B., Johnson, Z. I., Tolonen, A. C., Rohwer, F., & Chisholm, S. W. (2004). Transfer of photosynthesis genes to and from Prochlorococcus viruses. *Proc Natl Acad Sci USA*, 101(30):11013–8.

Luzzatto, L. (1967). Regulation of the activity of glucose-6-phosphate dehydrogenase by NADP+ and NADPH. *Biochim Biophys Acta*, 146(1):18–25.

Maciejewska, U. & Kacperska, A. (1987). Changes in the level of oxidized and reduced pyridine nucleotides during cold acclimation of winter rape plants. *Physiol Plantarum*, 69:687–691.

Mackey, K. R. M., Paytan, A., Grossman, A. R., & Bailey, S. (2008). A photosynthetic strategy for coping in a high-light, low-nutrient environment. *Limnol Oceanogr*, 53(3):900–13.

Maeda, N., Fan, H., & Yoshikai, Y. (2008). Oncogenesis by retroviruses: old and new paradigms. *Rev Med Virol*, 18(6):387–405.

Mann, N. H., Clokie, M. R. J., Millard, A., Cook, A., Wilson, W. H., Wheatley, P. J., Letarov, A., & Krisch, H. M. (2005). The genome of S-PM2, a "photosynthetic" T4-type bacteriophage that infects marine Synechococcus strains. *J Bacteriol*, 187(9):3188–200.

Mann, N. H., Cook, A., Millard, A., Bailey, S., & Clokie, M. (2003). Marine ecosystems: bacterial photosynthesis genes in a virus. *Nature*, 424(6950):741.

Markowitz, V. M., Korzeniewski, F., Palaniappan, K., Szeto, E., Werner, G., Padki, A., Zhao, X., Dubchak, I., Hugenholtz, P., Anderson, I., Lykidis, A., Mavromatis, K., Ivanova, N., & Kyrpides, N. C. (2006). The integrated microbial genomes (IMG) system. *Nucleic Acids Res*, 34:D344–8.

Marri, L., Trost, P., Trivelli, X., Gonnelli, L., Pupillo, P., & Sparla, F. (2008). Spontaneous assembly of photosynthetic supramolecular complexes as mediated by the intrinsically unstructured protein CP12. *J Biol Chem*, 283(4):1831–8.

Martiny, A. C., Coleman, M. L., & Chisholm, S. W. (2006). Phosphate acquisition genes in Prochlorococcus ecotypes: evidence for genome-wide adaptation. *Proc Natl Acad Sci USA*, 103(33):12552–7.

Mathews, C. K. (1977). Reproduction of large virulent bacteriophages. In: *Comprehensive Virology*, H. Fraenkel-Conrat & R. R. Wagner, ed., volume 7, pages 179–294. Plenum Press.

Mathews, C. K. & Kessin, R. H. (1967). Control of bacteriophage-induced enzyme synthesis in cells infected with a temperature-sensitive mutant. *J Virol*, 1(1):92–6.

McCoy, A. J., Grosse-Kunstleve, R. W., Storoni, L. C., & Read, R. J. (2005). Likelihood-enhanced fast translation functions. *Acta Crystallogr D Biol Crystallogr*, 61(Pt 4):458–64.

Millard, A., Clokie, M. R. J., Shub, D. A., & Mann, N. H. (2004). Genetic organization of the psbAD region in phages infecting marine Synechococcus strains. *Proc Natl Acad Sci USA*, 101(30):11007–12.

Millard, A. D., Zwirglmaier, K., Downey, M. J., Mann, N. H., & Scanlan, D. J. (2009). Comparative genomics of marine cyanomyoviruses reveals the widespread occurrence of Synechococcus host genes localized to a hyperplastic region: implications for mechanisms of cyanophage evolution. *Environ Microbiol*, 11(9):2370–87.

Miller, E. S., Heidelberg, J. F., Eisen, J. A., Nelson, W. C., Durkin, A. S., Ciecko, A., Feldblyum, T. V., White, O., Paulsen, I. T., Nierman, W. C., Lee, J., Szczypinski, B., & Fraser, C. M. (2003a). Complete genome sequence of the broad-host-range vibriophage KVP40: comparative genomics of a T4-related bacteriophage. *J Bacteriol*, 185(17):5220–33.

Miller, E. S., Kutter, E., Mosig, G., Arisaka, F., Kunisawa, T., & Rüger, W. (2003b). Bacteriophage T4 genome. *Microbiol Mol Biol Rev*, 67(1):86–156.

Minor, W., Cymborowski, M., Otwinowski, Z., & Chruszcz, M. (2006). HKL-3000: the integration of data reduction and structure solution–from diffraction images to an initial model in minutes. *Acta Crystallogr D Biol Crystallogr*, 62(Pt 8):859–66.

Miosga, T., Schaaff-Gerstenschlager, I., Franken, E., & Zimmermann, F. K. (1993). Lysine144 is essential for the catalytic activity of Saccharomyces cerevisiae transaldolase. *Yeast*, 9(11):1241–9.

Moore, L., Coe, A., Zinser, E., Saito, M., Sullivan, M., Lindell, D., Frois-Moniz, K., Waterbury, J., & Chisholm, S. (2007). Culturing the marine cyanobacterium Prochlorococcus. *Limnol Oceanogr Methods*, 5:353–362.

Murshudov, G. N., Vagin, A. A., & Dodson, E. J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr*, 53(Pt 3):240–55.

Padan, E. & Shilo, M. (1973). Cyanophages-viruses attacking blue-green algae. *Bacteriol Rev*, 37(3):343–70.

Page, R. D. M. (2002). Visualizing phylogenetic trees using TreeView. *Curr Protoc Bioinformatics*, Chapter 6:Unit 6.2.

Partensky, F., Hess, W. R., & Vaulot, D. (1999). Prochlorococcus, a marine photosynthetic prokaryote of global significance. *Microbiol Mol Biol Rev*, 63(1):106–27.

Paul, J. H., Sullivan, M. B., Segall, A. M., & Rohwer, F. (2002). Marine phage genomics. *Comp Biochem Physiol B Biochem Mol Biol*, 133(4):463–76.

Pedersen, A., Karlsson, G. B., & Rydström, J. (2008). Proton-translocating transhydrogenase: an update of unsolved and controversial issues. *J Bioenerg Biomembr*, 40(5):463–73.

Petrov, V. M., Nolan, J. M., Bertrand, C., Levy, D., Desplats, C., Krisch, H. M., & Karam, J. D. (2006). Plasticity of the gene functions for DNA replication in the T4-like phages. *J Mol Biol*, 361(1):46–68.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera–a visualization system for exploratory research and analysis. *J Comput Chem*, 25(13):1605–12.

Pfaffl, M. W. (2001). A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res*, 29(9):e45.

Pietro, A. S. & Lang, H. M. (1958). Photosynthetic pyridine nucleotide reductase. I. Partial purification and properties of the enzyme from spinach. *J Biol Chem*, 231(1):211–29.

Pohlmeyer, K., Paap, B. K., Soll, J., & Wedel, N. (1996). CP12: a small nuclear-encoded chloroplast protein provides novel insights into higher-plant GAPDH evolution. *Plant Mol Biol*, 32(5):969–78.

Pope, W. H., Weigele, P. R., Chang, J., Pedulla, M. L., Ford, M. E., Houtz, J. M., Jiang, W., Chiu, W., Hatfull, G. F., Hendrix, R. W., & King, J. (2007). Genome sequence, structural proteins, and capsid organization of the cyanophage Syn5: a "horned" bacteriophage of marine Synechococcus. *J Mol Biol*, 368(4):966–81.

Prasad, G. S., Sridhar, V., Yamaguchi, M., Hatefi, Y., & Stout, C. D. (1999). Crystal structure of transhydrogenase domain III at 1.2 A resolution. *Nat Struct Biol*, 6(12):1126–31.

Rahoutei, J., Garcia-Luque, I., & Baron, M. (2000). Inhibition of photosynthesis by viral infection: effect on PSII structure and function. *Physiol Plant*, 110:286–292.

Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N. A., Arellano, A., Coleman, M., Hauser, L., Hess, W. R., Johnson, Z. I., Land, M., Lindell, D., Post, A. F., Regala, W., Shah, M., Shaw, S. L., Steglich, C., Sullivan, M. B., Ting, C. S., Tolonen, A., Webb, E. A., Zinser, E. R., & Chisholm, S. W. (2003). Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. *Nature*, 424(6952):1042–7.

Rodionov, D. A., Vitreschak, A. G., Mironov, A. A., & Gelfand, M. S. (2003). Comparative genomics of the vitamin B12 metabolism and regulation in prokaryotes. *J Biol Chem*, 278(42):41148–59.

Roucourt, B. & Lavigne, R. (2009). The role of interactions between phage and bacterial proteins within the infected cell: a diverse and puzzling interactome. *Environ Microbiol*, 11(11):2789–805.

Rozen, S. & Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol*, 132:365–86.

Rusch, D., Halpern, A., Sutton, G., Heidelberg, K., Williamson, S., Yooseph, S., Wu, D., Eisen, J., Hoffman, J., Remington, K., Beeson, K., Tran, B., Smith, H., Baden-Tillson, H., Stewart, C., Thorpe, J., Freeman, J., Andrews-Pfannkoch, C., Venter, J., Li, K., Kravitz, S., Heidelberg, J., Utterback, T., Rogers, Y., Falcón, L., Souza, V., Bonilla-Rosso, G., Eguiarte, L., Karl, D., Sathyendranath, S., Platt, T., Bermingham, E., Gallardo, V., Tamayo-Castillo, G., Ferrari, M., Strausberg, R., Nealson, K., Friedman, R., Frazier, M., & Venter, J. (2007). The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biol*, 5(3):e77.

Sakiyama, S. & Buchanan, J. M. (1971). In vitro synthesis of deoxynucleotide kinase programmed by bacteriophage T4-RNA. *Proc Natl Acad Sci USA*, 68(6):1376–80.

Samland, A. K. & Sprenger, G. A. (2009). Transaldolase: from biochemistry to human disease. *Int J Biochem Cell Biol*, 41(7):1482–94.

Sandaa, R.-A., Gómez-Consarnau, L., Pinhassi, J., Riemann, L., Malits, A., Weinbauer, M. G., Gasol, J. M., & Thingstad, T. F. (2009). Viral control of bacterial biodiversity–evidence from a nutrient-enriched marine mesocosm experiment. *Environ Microbiol*, 11(10):2585–97.

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, C. A., Hutchison, C. A., Slocombe, P. M., & Smith, M. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, 265(5596):687–95.

Scanlan, D. & West, N. (2002). Molecular ecology of the marine cyanobacterial genera Prochlorococcus and Synechococcus. *FEMS Microbiol Ecol*, 40:1–12.

Scanlan, D. J., Ostrowski, M., Mazard, S., Dufresne, A., Garczarek, L., Hess, W. R., Post, A. F., Hagemann, M., Paulsen, I., & Partensky, F. (2009). Ecological genomics of marine picocyanobacteria. *Microbiol Mol Biol Rev*, 73(2):249–99.

Schneider, S., Sandalova, T., Schneider, G., Sprenger, G. A., & Samland, A. K. (2008). Replacement of a phenylalanine by a tyrosine in the active site confers fructose-6-phosphate aldolase activity to the transaldolase of Escherichia coli and human origin. *J Biol Chem*, 283(44):30064–72.

Schörken, U., Thorell, S., Schürmann, M., Jia, J., Sprenger, G. A., & Schneider, G. (2001). Identification of catalytically important residues in the active site of Escherichia coli transaldolase. *Eur J Biochem*, 268(8):2408–15.

Schürmann, M. & Sprenger, G. A. (2001). Fructose-6-phosphate aldolase is a novel class I aldolase from Escherichia coli and is related to a novel group of bacterial transaldolases. *J Biol Chem*, 276(14):11055–61.

Schwede, T., Kopp, J., Guex, N., & Peitsch, M. C. (2003). SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res*, 31(13):3381–5.

Seshadri, R., Kravitz, S. A., Smarr, L., Gilna, P., & Frazier, M. (2007). CAMERA: a community resource for metagenomics. *PLoS Biol*, 5(3):e75.

Sharon, I., Alperovitch, A., Rohwer, F., Haynes, M., Glaser, F., Atamna-Ismaeel, N., Pinter, R. Y., Partensky, F., Koonin, E. V., Wolf, Y. I., Nelson, N., & Béjà, O. (2009). Photosystem I gene cassettes are present in marine virus genomes. *Nature*, 461(7261):258–62.

Sharon, I., Tzahor, S., Williamson, S., Shmoish, M., Man-Aharonovich, D., Rusch, D. B., Yooseph, S., Zeidner, G., Golden, S. S., Mackey, S. R., Adir, N., Weingart, U., Horn, D., Venter, J. C., Mandel-Gutfreund, Y., & Béjà, O. (2007). Viral photosynthetic reaction center genes and transcripts in the marine environment. *ISME J*, 1(6):492–501.

Sherman, L. A. (1976). Infection of Synechococcus cedrorum by the cyanophage AS-1M. III. Cellular metabolism and phage development. *Virology*, 71(1):199–206.

Soderberg, T. & Alver, R. C. (2004). Transaldolase of Methanocaldococcus jannaschii. *Archaea*, 1(4):255–62.

Sprenger, G. A., Schörken, U., Sprenger, G., & Sahm, H. (1995). Transaldolase B of Escherichia coli K-12: cloning of its gene, talB, and characterization of the enzyme from recombinant strains. *J Bacteriol*, 177(20):5930–6.

Stanier, R. Y. & Cohen-Bazire, G. (1977). Phototrophic prokaryotes: the cyanobacteria. *Annu Rev Microbiol*, 31:225–74.

Stöckel, J., Welsh, E. A., Liberton, M., Kunnvakkam, R., Aurora, R., & Pakrasi, H. B. (2008). Global transcriptomic analysis of Cyanothece 51142 reveals robust diurnal oscillation of central metabolic processes. *Proc Natl Acad Sci USA*, 105(16):6156–61.

Stubbe, J., Ge, J., & Yee, C. S. (2001). The evolution of ribonucleotide reduction revisited. *Trends Biochem Sci*, 26(2):93–9.

Sullivan, M. B., Coleman, M. L., Weigele, P., Rohwer, F., & Chisholm, S. W. (2005). Three Prochlorococcus cyanophage genomes: signature features and ecological interpretations. *PLoS Biol*, 3(5):e144.

Sullivan, M. B., Huang, K. H., Ignacio-Espinoza, J. C., Berlin, A., Kelly, L., Weigele, P. R., DeFrancesco, A. S., Kern, S. E., Thompson, L. R., Young, S., Lee, W., Weiand, M., Fu, R., Krastins, B., Chase, M., Sarracino, D., Osburne, M. S., Henn, M. R., & Chisholm, S. W. (in press). Genomic analysis of oceanic cyanobacterial myoviruses compared to T4-like myoviruses from diverse hosts and environments. *Environ Microbiol*.

Sullivan, M. B., Krastins, B., Hughes, J. L., Kelly, L., Chase, M., Sarracino, D., & Chisholm, S. W. (2009). The genome and structural proteome of an ocean siphovirus: a new window into the cyanobacterial 'mobilome'. *Environ Microbiol*, 11(11):2935–51.

Sullivan, M. B., Lindell, D., Lee, J. A., Thompson, L. R., Bielawski, J. P., & Chisholm, S. W. (2006). Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biol*, 4(8):e234.

Sullivan, M. B., Waterbury, J. B., & Chisholm, S. W. (2003). Cyanophages infecting the oceanic cyanobacterium Prochlorococcus. *Nature*, 424(6952):1047–51.

Sundaram, S., Karakaya, H., Scanlan, D. J., & Mann, N. H. (1998). Multiple oligomeric forms of glucose-6-phosphate dehydrogenase in cyanobacteria and the role of OpcA in the assembly process. *Microbiology*, 144:1549–56.

Suttle, C. A. (2005). Viruses in the sea. *Nature*, 437(7057):356–61.

Suttle, C. A. & Chan, A. M. (1994). Dynamics and distribution of cyanophages and their effect on marine Synechococcus spp. *Appl Environ Microbiol*, 60(9):3167–3174.

Tamoi, M., Miyazaki, T., Fukamizo, T., & Shigeoka, S. (2005). The Calvin cycle in cyanobacteria is regulated by CP12 via the NAD(H)/NADP(H) ratio under light/dark conditions. *Plant J*, 42(4):504–13.

Thingstad, T. F., Krom, M. D., Mantoura, R. F. C., Flaten, G. A. F., Groom, S., Herut, B., Kress, N., Law, C. S., Pasternak, A., Pitta, P., Psarra, S., Rassoulzadegan, F., Tanaka, T., Tselepides, A., Wassmann, P., Woodward, E. M. S., Riser, C. W., Zodiatis, G., & Zohary, T. (2005). Nature of phosphorus limitation in the ultraoligotrophic eastern Mediterranean. *Science*, 309(5737):1068–71.

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–80.

Thorell, S., Gergely, P., Banki, K., Perl, A., & Schneider, G. (2000). The three-dimensional structure of human transaldolase. *FEBS Lett*, 475(3):205–8.

Thorell, S., Schürmann, M., Sprenger, G. A., & Schneider, G. (2002). Crystal structure of decameric fructose-6-phosphate aldolase from Escherichia coli reveals inter-subunit helix swapping as a structural basis for assembly differences in the transaldolase family. *J Mol Biol*, 319(1):161–71.

Tillett, H. E. (1987). Most probable numbers of organisms: revised tables for the multiple tube method. *Epidemiol Infect*, 99(2):471–6.

Toepel, J., Welsh, E., Summerfield, T. C., Pakrasi, H. B., & Sherman, L. A. (2008). Differential transcriptional analysis of the cyanobacterium Cyanothece sp. strain ATCC 51142 during light-dark and continuous-light growth. *J Bacteriol*, 190(11):3904–13.

Vaulot, D., Marie, D., Olson, R., & Chisholm, S. (1995). Growth of Prochlorococcus, a photosynthetic prokaryote, in the equatorial Pacific Ocean. *Science*, 268(5216):1480–1482.

Waard, A. D., Paul, A. V., & Lehman, I. R. (1965). The structural gene for deoxyribonucleic acid polymerase in bacteriophages T4 and T5. *Proc Natl Acad Sci USA*, 54(4):1241–8.

Waldbauer, J. R. (2009). *Molecular biogeochemistry of modern and ancient marine microbes*. PhD Thesis, Joint Program in Chemical Oceanography, Massachusetts Institute of Technology and Woods Hole Oceanographic Institution.

Waterbury, J. & Valois, F. (1993). Resistance to co-occurring phages enables marine Synechococcus communities to coexist with cyanophages abundant in seawater. *Appl Environ Microbiol*, 59(10):3393–3399.

Waterbury, J. & Willey, J. (1988). Isolation and growth of marine planktonic cyanobacteria. *Methods Enzymol*, 167:100–105.

Wedel, N. & Soll, J. (1998). Evolutionary conserved light regulation of Calvin cycle activity by NADPH-mediated reversible phosphoribulokinase/CP12/glyceraldehyde-3-phosphate dehydrogenase complex dissociation. *Proc Natl Acad Sci USA*, 95(16):9699–704.

Wedel, N., Soll, J., & Paap, B. K. (1997). CP12 provides a new mode of light regulation of Calvin cycle activity in higher plants. *Proc Natl Acad Sci USA*, 94(19):10479–84.

Weigele, P. R., Pope, W. H., Pedulla, M. L., Houtz, J. M., Smith, A. L., Conway, J. F., King, J., Hatfull, G. F., Lawrence, J. G., & Hendrix, R. W. (2007). Genomic and structural analysis of Syn9, a cyanophage infecting marine Prochlorococcus and Synechococcus. *Environ Microbiol*, 9(7):1675–95.

Williamson, S. J., Rusch, D. B., Yooseph, S., Halpern, A. L., Heidelberg, K. B., Glass, J. I., Andrews-Pfannkoch, C., Fadrosh, D., Miller, C. S., Sutton, G., Frazier, M., & Venter, J. C. (2008). The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. *PLoS ONE*, 3(1):e1456.

Wood, T. (1986a). Distribution of the pentose phosphate pathway in living organisms. *Cell Biochem Funct*, 4(4):235–40.

Wood, T. (1986b). Physiological functions of the pentose phosphate pathway. *Cell Biochem Funct*, 4(4):241–7.

Wu, J., Sunda, W., Boyle, E. A., & Karl, D. M. (2000). Phosphate depletion in the western North Atlantic Ocean. *Science*, 289(5480):759–62.

Yasuda, S. & Sekiguchi, M. (1970). T4 endonuclease involved in repair of DNA. *Proc Natl Acad Sci USA*, 67(4):1839–45.

Yerrapragada, S., Siefert, J. L., & Fox, G. E. (2009). Horizontal gene transfer in cyanobacterial signature genes. *Methods Mol Biol*, 532:339–66.

Zeidner, G., Bielawski, J. P., Shmoish, M., Scanlan, D. J., Sabehi, G., & Béjà, O. (2005). Potential photosynthesis gene recombination between Prochlorococcus and Synechococcus via viral intermediates. *Environ Microbiol*, 7(10):1505–13.

Zhang, R. G., Skarina, T., Katz, J. E., Beasley, S., Khachatryan, A., Vyas, S., Arrowsmith, C. H., Clarke, S., Edwards, A., Joachimiak, A., & Savchenko, A. (2001). Structure of Thermotoga maritima stationary phase survival protein SurE: a novel acid phosphatase. *Structure*, 9(11):1095–106.

Zinser, E. R., Coe, A., Johnson, Z. I., Martiny, A. C., Fuller, N. J., Scanlan, D. J., & Chisholm, S. W. (2006). Prochlorococcus ecotype abundances in the North Atlantic Ocean as revealed by an improved quantitative PCR method. *Appl Environ Microbiol*, 72(1):723–32.

Zinser, E. R., Lindell, D., Johnson, Z. I., Futschik, M. E., Steglich, C., Coleman, M. L., Wright, M. A., Rector, T., Steen, R., McNulty, N., Thompson, L. R., & Chisholm, S. W. (2009). Choreography of the transcriptome, photophysiology, and cell cycle of a minimal photoautotroph, Prochlorococcus. *PLoS ONE*, 4(4):e5135.

Zwirglmaier, K., Jardillier, L., Ostrowski, M., Mazard, S., Garczarek, L., Vaulot, D., Not, F., Massana, R., Ulloa, O., & Scanlan, D. J. (2008). Global phylogeography of marine Synechococcus and Prochlorococcus reveals a distinct partitioning of lineages among oceanic biomes. *Environ Microbiol*, 10(1):147–61.